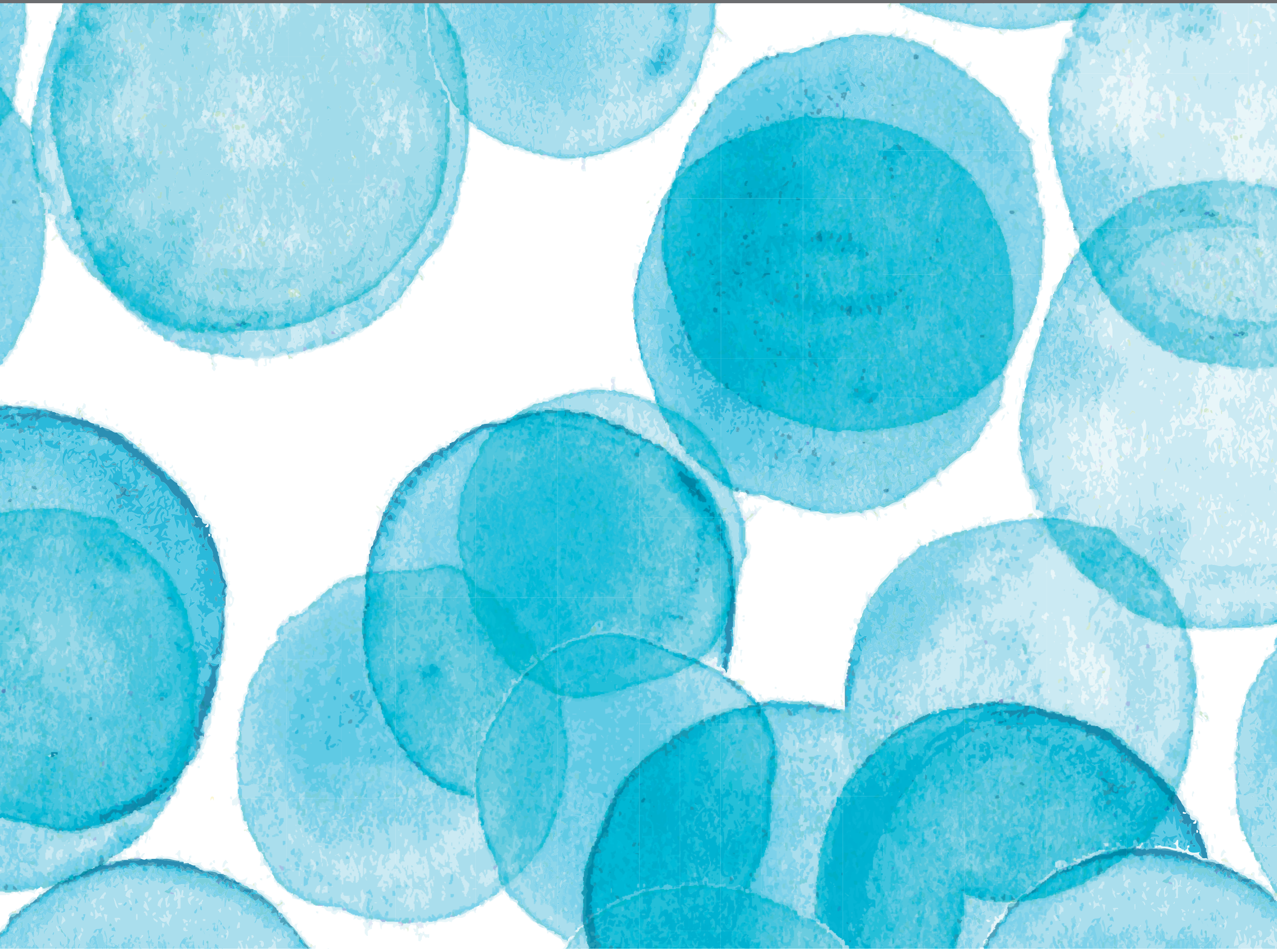
VIRUSES, GENETIC EXCHANGE, AND THE TREE OF LIFE

EDITED BY: Arshan Nasir, Gustavo Caetano-Anollés and
Jean-Michel Claverie

PUBLISHED IN: Frontiers in Microbiology and Frontiers in Genetics



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-368-5
DOI 10.3389/978-2-88963-368-5

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: researchtopics@frontiersin.org

VIRUSES, GENETIC EXCHANGE, AND THE TREE OF LIFE

Topic Editors:
Arshan Nasir, COMSATS University Islamabad, Pakistan; Los Alamos National Laboratory (DOE), United States
Gustavo Caetano-Anollés, University of Illinois at Urbana-Champaign, United States
Jean-Michel Claverie, Aix-Marseille Université, France

Citation: Nasir, A., Caetano-Anollés, G., Claverie, J.-M., eds. (2020). Viruses, Genetic Exchange, and the Tree of Life. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-368-5

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 29 November 2019
doi: 10.3389/fmicb.2019.02782

Editorial: Viruses, Genetic Exchange, and the Tree of Life

Arshan Nasir[1,2]*, *Gustavo Caetano-Anollés*[3]* and *Jean-Michel Claverie*[4]*

[1] Department of Biosciences, COMSATS University Islamabad, Islamabad, Pakistan, [2] Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, United States, [3] Evolutionary Bioinformatics Laboratory, Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, United States, [4] Aix Marseille University, CNRS, IGS, Structural and Genomic Information Laboratory (UMR7256), Mediterranean Institute of Microbiology (FR3479), Marseille, France

Keywords: viruses, tree of life, evolution, horizontal gene transfer, ORFans, origin of nucleus

Editorial on the Research Topic

Viruses, Genetic Exchange, and the Tree of Life

We live in exciting times for microbiology research. The significant technological and scientific advancements in the past decades have now enabled scientists to pursue discovery and description of novel microbial and viral linages from previously uncharted Earth habitats. Two such discoveries are particularly exciting and noteworthy in this regard. First, the discovery of the first "giant virus," *Acanthamoeba polyphaga mimivirus*, in 2003, and several others thereafter, posed intriguing questions regarding virus origins, evolution, classification, and their place in the "tree of life." Second, the discoveries of "Lokiarchaeota" and several other closely-related archaeal members that encode several eukaryote-specific proteins challenged the three-domain canonical structure of the tree of life. These discoveries have reopened debates on central questions in evolutionary biology research such as the origin of viruses, the origin of eukaryotes, evolutionary relationship between Archaea and Eukarya, and the structure and topology of the tree of life. In this Research Topic, we received a broad range of contributions addressing these and other related questions.

Moelling and Broecker discussed the "virus first" model for the evolution of life on Earth. They provided several examples of virus diversity and abundance in a range of Earth environments and in the mammalian genomes. According to their view, ribozymes and viroids could have started early evolution albeit they also acknowledged competing alternatives such as the "proteins first" and the "metabolism first" scenarios of origins of life. In a separate contribution from the same authors, they examined the possibility that complex antiviral defense strategies and immune systems in cellular organisms could have evolved from viruses and transposable elements. The authors thus highlighted the crucial roles viruses could have played during the evolution of cells.

Ramisetty and Sudhakari discussed why the selection of "grounded prophages" may be favored by bacterial cells. Grounded prophages are unable to excise from bacterial genomes due to mutations in either recombinase gene or genes encoding attachment sites. Lysogens with grounded prophages are protected from specific phage infections and the future activation of lytic cycles. These grounded prophages can also serve as hotspots or buffer zones where genes encoding antibiotic resistance and virulence may integrate. Their work thus highlights another important contribution of viruses to the evolution of cells that contradicts the view of viruses as mere cellular pathogens.

Legendre et al. highlighted an often-ignored feature of viral genomes, the existence of many protein-coding genes with no detectable homologs. These virus-encoded ORFans constitute the large majority of genes of giant viruses but are mostly ignored when discussing models of virus evolution, while their origin remains a mystery. The authors provided evidence that ORFans

OPEN ACCESS

Edited by:
Nikos Kyrpides,
Lawrence Berkeley National Laboratory, United States

Reviewed by:
Simon Roux,
Lawrence Berkeley National Laboratory, United States

****Correspondence:***
Arshan Nasir
arshan.nasir@gmail.com
Gustavo Caetano-Anollés
gca@illinois.edu
Jean-Michel Claverie
claverie@igs.cnrs-mrs.fr

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

Received: *22 August 2019*
Accepted: *14 November 2019*
Published: *29 November 2019*

Citation:
Nasir A, Caetano-Anollés G and Claverie J-M (2019) Editorial: Viruses, Genetic Exchange, and the Tree of Life. Front. Microbiol. 10:2782. doi: 10.3389/fmicb.2019.02782

in pandoraviruses, the largest viruses known to date, can be unique even among closely related members of the same virus family. These ORFans likely originate from intergenic regions and suggest that the pandoravirus pan-genome is open. These findings indicate that the genomes of viruses are incredibly dynamic in their gene creation capabilities, a notion hardly acknowledged in standard virus evolutionary scenarios where all genes are bound to have an ancestor.

The presence of nucleus is a defining distinction between eukaryotes and prokaryotes. While nucleus-like compartmentation has been described or proposed in planctomycetes and jumbophage-infected bacteria, it is unclear how these compartmentations are linked to the origin of nucleus in eukaryotes. Hendrickson and Poole presented several analogies of nucleus-like compartmentation outside eukaryotes and discussed three possible explanations for the emergence of compartmentation in cells: physical protection, crosstalk avoidance, and non-adaptive origins.

Harrison et al. explored the diversity of ribonucleotide reductases (RNRs) from marine viroplankton, ancient enzymes that reduce ribonucleotides to deoxyribonucleotides and are prominent in viral genomes. They found cyanophage Class II RNR enzymes were misannotated and were actually part of the large oxygen-dependent Class I clade. Since the marine *Synechococcus* and *Prochlorococcus* hosts carry only Class II enzymes, the finding confirms that the viroplankton RNRs are not host derived nor dependent on the adenosylcobalamin (B12) cofactor of class II enzymes. This is an important clarification for connecting genomic information and phenotypic traits in the context of viral ecology.

Wang et al. analyzed the genomes of 21 Coxsackievirus A4 isolates from hand, foot and mouth disease cases in children from the Shandong province of China. Phylogenetic analysis of the VP1 gene, when benchmarked to a large Coxsackievirus collection, revealed that genotypes C, D1, and D2 identified in the early 2000s have been taken over by the D2 genotype in China. Calculation of substitution rates revealed ongoing virus evolution and several dynamic fluctuations in the history of the virus isolates. The study is important for the molecular epidemiological characterization of A4, which has been understudied because of the paucity of genomic data.

Flynn and Moreau engaged in a comparative host-centered exploration of endogenous viral elements (EVE) in ants. Using a comprehensive bioinformatic pipeline, they screened all 19 published ant genomes for EVEs and assessed their phylogenetic relationships to closely related exogenous viruses. EVEs similar to proteins from single stranded RNA viruses, viral glycoproteins, and retro-virus-derived proteins were widely present in ant genomes, suggesting tendencies to endogenize. EVEs therefore mimic the diversity of viral lineages.

Ongrádi et al. isolated an adenovirus from a domestic cat and found it was related to human adenovirus. Their careful molecular, biological, and phylogenetic characterizations highlight important information with the potential to re-define the adenovirus research field. It also prompts a broader analysis of adenoviruses of many different types. Further epidemiological and pathomechanistic studies can help understand the molecular mechanisms of virus host jumps, which could be engineered to combat both feline and human AIDS.

Flodman et al. performed an extensive analysis of interaction between commercially available restriction enzymes and bacteriophages containing modified nucleotides. They evaluated restriction resistance of phages M6, Vil, and phi W-14 against 200 commercially available enzymes. This work has enormous value for genetic engineering and understanding interaction of proteins with virus-modified DNA.

Sun et al. showed evidence that polysaccharides isolated from the leaves of the *Aloe vera* plant, known to mitigate virus infection, inhibit the replication of a H_1N_1 subtype of influenza virus at the time of virus adsorption. The anti-influenza activities were for the first time explored *in vitro* in cell cultures and in a mouse model. This study opens the door to the development of novel anti-influenza drugs.

Finally, Zhou et al. presented a novel mismatch-tolerant loop-mediated isothermal amplification (LAMP) method with an efficiency high enough to be applied to the detection of viruses. The methodology tolerates mismatches in primers and templates. They used this new methodology to detect variants of antigenically-distinct serotypes of the dengue virus.

The Research Topic is thus a collection of ideas ranging from our basic understanding of virus origins and evolution to new approaches of virus identification and treatment. We hope this resource will be a useful reference for future studies in basic and applied virology.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

ACKNOWLEDGMENTS

This work was supported by Oppenheimer Fellowship (20180751PRD3), Los Alamos National Laboratory to AN. Grants from the National Science Foundation (OISE-1132791) and the National Institute of Food and Agriculture of the United States Department of Agriculture (ILLU-802-909 and ILLU-483-625) to GC-A.

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.

Aloe Polysaccharides Inhibit Influenza A Virus Infection—A Promising Natural Anti-flu Drug

Zhenhong Sun[1]†, Cuilian Yu[2]†, Wei Wang[3], Guangfu Yu[1], Tingting Zhang[1], Lin Zhang[2], Jiguo Zhang[1]* and Kai Wei[2]*

[1] School of Basic Medical Sciences, Taishan Medical University, Tai'an, China, [2] College of Animal Science and Technology, Shandong Agricultural University, Tai'an, China, [3] Guangdong Winsun Bio-pharmaceutical Co., Ltd., Guangzhou, China

Influenza A virus causes periodic outbreaks and seriously threatens human health. The drug-resistant mutants have shown an epidemic trend because of the abuse of chemical drugs. *Aloe* polysaccharides (APS) extracted from *Aloe vera* leaves have evident effects on the therapy of virus infection. However, the activity of APS in anti-influenza virus has yet to be investigated. Here, we refined polysaccharides from *A. vera* leaf. *In vitro* test revealed that APS could inhibit the replication of a H1N1 subtype influenza virus, and the most obvious inhibitory effect was observed in the viral adsorption period. Transmission electron microscopy indicated that APS directly interacted with influenza virus particles. Experiments on PR8 (H1N1) virus infection in mice demonstrated that APS considerably ameliorated the clinical symptoms and the lung damage of the infected mice, and significantly reduced the virus loads and mortality. Our findings provided a theoretical basis for the development of novel natural anti-influenza agents.

Keywords: *Aloe* polysaccharides, H1N1, antivirus, viral adsorption, transmission electron microscopy

OPEN ACCESS

Edited by:
*Gustavo Caetano-Anollés,
University of Illinois
at Urbana–Champaign, United States*

Reviewed by:
*Akatsuki Saito,
Osaka University, Japan
Martin Michaelis,
University of Kent, United Kingdom*

***Correspondence:**
*Jiguo Zhang
jiguotsmc@163.com
Kai Wei
weikaisdau@163.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: *10 July 2018*
Accepted: *12 September 2018*
Published: *27 September 2018*

Citation:
Sun Z, Yu C, Wang W, Yu G, Zhang T, Zhang L, Zhang J and Wei K (2018) Aloe Polysaccharides Inhibit Influenza A Virus Infection—A Promising Natural Anti-flu Drug. Front. Microbiol. 9:2338. doi: 10.3389/fmicb.2018.02338

INTRODUCTION

Influenza A virus (IAV) causes acute respiratory distress syndrome, annual epidemics, and occasional pandemics, which have claimed the lives of millions of individuals. Oseltamivir-resistant H1N1 influenza viruses are spread seasonally (Hibino et al., 2017; Leung et al., 2017), and swine-originated H1N1 viruses that have been transmitted to and spread among humans cause outbreaks worldwide (Liu et al., 2017; Ylipalosaari et al., 2017). Highly pathogenic H5N1 avian influenza virus infection continuously threatens poultry and human health (Ly et al., 2017). Some novel recombinant viruses, such as H7N9 and H5NX viruses, have also emerged and affected human health (Pan et al., 2016; Taubenberger and Morens, 2017). These results highlight the limitations of preventive and therapeutic measures against the influenza virus.

Vaccines and antiviral drugs are available for the control of influenza virus infections. Two kinds of antiviral drugs, namely, ion channel and neuraminidase inhibitors, are licensed for use against IAV. Human seasonal influenza vaccines, including H1N1 and H3N2, and some commercially available inactivated vaccines for avian influenza have been used (Nichol and Treanor, 2006; Swayne, 2009). Nonetheless, the global community is probably not well prepared for the next influenza pandemic because viruses have acquired resistance to currently available antiviral drugs. Excessive reliance on vaccines is also inadequate. Vaccines for human and avian influenza have not been accepted worldwide because of their varied protection levels, and traditional vaccines must be updated periodically to account for the antigenic drift and shift of circulating viruses (Lambert and Fauci, 2010). Moreover, the production of a vaccine for a newly emerging strain takes several weeks or months (Lambert and Fauci, 2010). During this time, a pandemic virus can spread globally. Therefore, novel therapeutic schedules are necessary to explore, develop, and control the dynamic and increasingly complicated ecology of circulating IAVs.

Medicinal plants have been widely used to treat various infectious and non-infectious ailments. Natural plant polysaccharides, which are polymeric carbohydrate molecules composed of long chains of monosaccharide units, have different biological activities, including anti-inflammatory activities, immunological regulation, oxidation resistance, and antiviral activities. Numerous studies have reported the inhibitory effects of plant polysaccharides, such as brown algal polysaccharides, *Auricularia auricula* polysaccharides, *Pinus massoniana* pollen polysaccharides, and *Acanthopanax sciadophylloides* polysaccharides, on the replication of human and animal viruses (Queiroz et al., 2008; Nguyen et al., 2012; Lee et al., 2015; Yu et al., 2017). In our study, we focused on *Aloe vera*, which is a perennial Liliaceae evergreen herbaceous plant widely grown in the tropical and subtropical regions. *A. vera* gel contains a large amount of bioactive ingredients, such as vitamins, amino acids, trace elements, polysaccharides, and anthraquinones; as such, it has antibacterial, anti-inflammatory, antioxidant, wound healing-promoting, and immunity-enhancing functions (Langmead et al., 2004; Yagi and Byung, 2015; Kumar and Tiku, 2016). *A. vera* L. can function as nutritional support for patients who are infected with human immunodeficiency virus in clinical trials and can simultaneously affect the viral capability of replication (Radha and Laxmipriya, 2015). *A. vera* also shows antiviral activity against herpes simplex virus, and *A. vera* gel extracts are conducive to the treatment of genital herpes in males (Syed et al., 1996; Zandi et al., 2007). Moreover, *A. vera* can be added to drugs for the treatment of high risk Human papillomavirus infection (Iljazovic et al., 2006). *Aloe* polysaccharides (APS) are active ingredients with a high content in *A. vera* gel. Acemannan polysaccharide is a representative acetylated mannan extracted from *A. vera* gel and has been approved by the US FDA for the treatment of AIDS in humans (Kahlon et al., 1991a,b). Gauntt et al. (2000) also reported that *Aloe* mannan can increase the titers of specific antibodies in mice infected with Coxsackievirus B3, thereby inducing antiviral effects. Therefore, we speculate that APS may have great potential to inhibit influenza virus infection.

This study aimed to explore the inhibitory effect and mechanism of APS on influenza virus infection. We extracted and purified APS from *A. vera* leaves, then we analyzed the monosaccharide composition. We next examined for the first time the anti-influenza activities of APS in a cell culture *in vitro*. Transmission electron microscopy (TEM) was further employed to examine the possible mechanism. The anti-influenza effect of APS was also evaluated in a mouse model. The performance of APS indicated its potential for the development of novel anti-influenza drugs.

MATERIALS AND METHODS

Ethics Statement

The animal procedures were approved by the Animal Care and Use Committee of Shandong Agricultural University (Permit number: 20010510), and performed according to the "Guidelines for Experimental Animals" of the Ministry of Science and Technology (Beijing, China).

Virus, Cells, and Polysaccharide

Influenza A/Puerto Rico/8/34 (PR8, subtype H1N1) was kept in our laboratory and titrated in MDCK cells by determining $TCID_{50}$. MDCK cells were maintained in Dulbecco's modified Eagle's medium (Gibco) containing 10% fetal bovine serum (Gibco) in 5% CO_2 at 37°C. APS was extracted from *A. vera* leaf through water extraction and ethanol precipitation, and the key steps refer to the method of our previous study (Sun et al., 2011). Briefly, dried *A. vera* leaves and water were mixed at the ratio of 1:10 and extracted at 85°C for 8 h. The extraction liquid was condensed by rotary evaporator. Then high speed centrifugation (12,000 rpm, 20 min) were performed to remove the impurities. Subsequently, ethanol was added into the concentrated solution to precipitate polysaccharides with the final concentration of 75%. After centrifugation and redissolution of polysaccharides, the ethanol precipitation process is repeated three times. Subsequently, the removal of protein was performed Sevag method [chloroform: *n*-butanol = 5:1 (V/V)], and this process was repeated five times. And finally via vacuum drying, the refined APS were obtained.

Determination of Molecular Weight (Mw) and Monosaccharide Composition

The Mw and monosaccharide composition of APS were determined by Qingdao Sci-tech Innovation Quality Testing Co., Ltd. (Qingdao, China). The Mw was measured by gel permeation chromatography (GPC) method. The monosaccharide composition was measured by Agilent 1200 high performance liquid chromatograph (HPLC, quaternary pump, autosampler, DAD detector, Agilent LC ChemStation). Briefly, APS was hydrolyzed with trifluoroacetic acid and derivatized with 1-phenyl-3-methyl-5-pyrazolone (PMP) and NaOH. Subsequently, HCl was added for neutralization, and monosaccharide extraction was performed thrice with chloroform, then the sample was tested. The monosaccharide standards include 10 kinds of monosaccharides, namely rhamnose, arabinose, galactose, xylose, glucose, mannose, ribosome, fucose, glucuronic acid, and galacturonic acid (Dr. Ehrenstorfer GmbH, Germany). Chromatographic column: Thermo C18 column (4.6 mm × 250 mm, 5.0 μm); Mobile phase: 0.1 mol/L of phosphate buffer solution (pH 7.0):acetonitrile = 82:18 (v/v); The flow rate: 1.0 mL/min; column temperature: 25°C; Sample quantity for 10 μL; wavelength: 245 nm.

Cytotoxicity Test of APS

MDCK cell monolayers were cultured in 96-well cell culture plates. Different APS concentrations (i.e., 20, 40, 80, 160, 320, and 640 μg/mL) were separately added to the cells. Three repeats were set for each concentration. After 72 h of incubation, the cell viability was determined with MTT assay, and the wavelength for reading was set at 490 nm (Mosmann, 1983).

Antiviral Activities of APS in MDCK Cells

The extracted APS was diluted to 40, 80, and 160 μg/mL with a maintenance medium (MM). The MDCK cell monolayers grown in 24-well plates were covered with 1.0 mL of MM containing the corresponding APS concentrations and infected with PR8 virus [multiplicity of infection (MOI) = 1]. After 1 h of viral adsorption at 37°C, the supernatants in the wells were replaced with the fresh MM containing the corresponding APS concentrations. Then, the growth dynamics of the virus were detected, and the viral titers of the cell culture supernatant collected every 12 h were determined.

The MDCK cells grown in 24-well plates were inoculated with PR8 virus (MOI = 1) and treated with APS at a final concentration of 40 μg/mL at the following artificially divided viral infection phases to investigate the action phase of APS in the viral life cycle:

Before Adsorption (BA)

H1N1 virus was preincubated with APS for 1 h at 4°C and subsequently used for infection. The MDCK cells were infected with the complex of H1N1 virus and polysaccharide for another 1 h at 37°C. After the supernatant was removed, the cells were washed twice and recovered with pure MM.

Adsorption (Ad)

MDCK cells were exposed to MM containing the virus and APS for 1 h at 37°C. After the supernatant was removed, the cells were washed twice and recovered with pure MM.

After Adsorption (AA)

MDCK cells were infected with H1N1 virus in the absence of APS. After viral adsorption occurred for 1 h at 37°C, the non-adherent viruses were removed. The cells were washed twice and subsequently incubated with MM containing APS.

All of the supernatants in the cell culture wells were collected at each 24 h interval and titrated through $TCID_{50}$ assay. Indirect immunofluorescence assay was performed to identify the virus-infected cells by using the mouse anti-influenza NP monoclonal antibody (Abcam, Cambridge, United Kingdom) at 4 days post-infection. The MDCK cells infected with H1N1 virus in the absence of APS in the entire assay served as the control.

TEM Assay

The MDCK cell monolayers were infected with PR8 virus suspension (MOI = 1) containing APS (40 μg/mL). The monolayers were scraped at 4 h post-infection and successively fixed in 2.5% glutaraldehyde and 1% osmium tetroxide at 4°C. The cell samples were dehydrated in a graded acetone series prior to infiltration and embedding. Ultrathin (50 nm, LKB-V) longitudinal sections were prepared after the samples were located in the semithin section, stained with uranyl acetate and lead citrate, and examined under a JEOL-1200EX electron microscope (JEOL, Japan). The infected cells without APS treatment serve as the control. Additionally, the PR8 virus suspension was concentrated through sucrose density gradient centrifugation. The isovolumetric virus suspension (0.55 mg/mL) and APS (40 μg/mL) were mixed and incubated for 1 h at 4°C. The mixture was subsequently placed on carbon-coated grids and negatively stained with 0.01 mL of 2% phosphotungstic acid for

TABLE 1 | The Mw determination of APS.

Mw (KDa)	Polydispersity (PD):	1.718
	Number-average Mw (Mn):	5354
	Weight-average Mw (Mw) :	9198
	Z-average Mw (Mz) :	16817
Mw distribution	<100, %	76.823
	100∼500, %	14.213
	500∼1000, %	5.624
	1000∼5000, %	1.607
	5000∼10000, %	0.959
	>10000, %	0.774

TABLE 2 | The relative content of monosaccharide composition in APS.

Numerical order	Monosaccharide type	Monosaccharide content (%)
1	Mannose	5.85
2	Ribose	1.50
3	Rhamnose	2.20
4	Glucuronic acid	5.78
5	Galacturonic acid	7.84
6	Glucose	7.57
7	Galactose	52.88
8	Xylose	0 (not detected)
9	Arabinose	10.07
10	Fucose	6.30

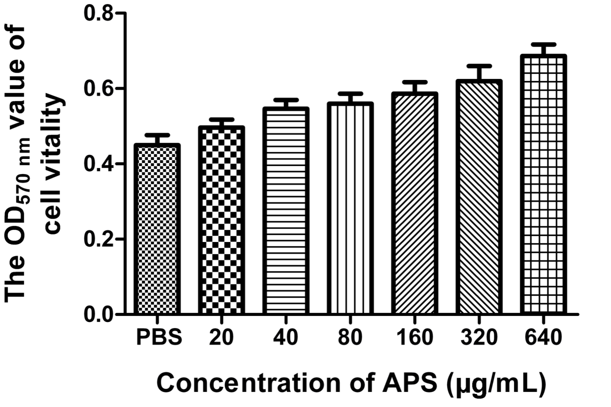


FIGURE 1 | Influence of APS on cell activity. MDCK cells were cultured with the various concentrations of APS (20, 40, 80, 160, 320, and 640 μg/mL) or PBS for 72 h. The cell viability was determined by MTT assay.

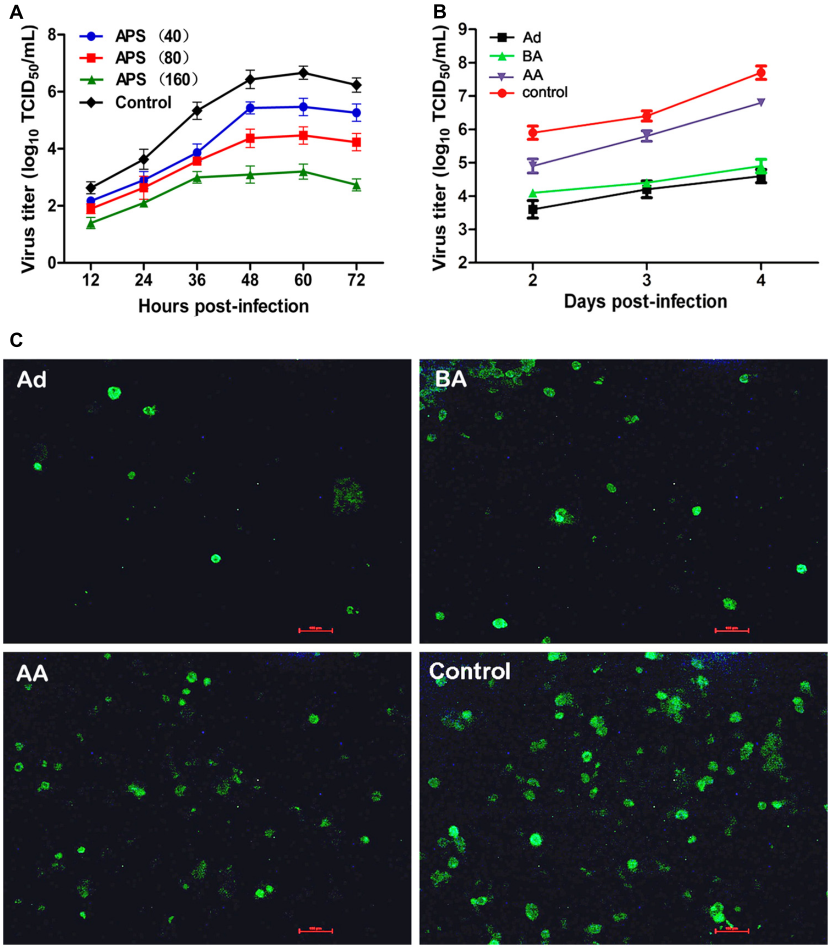


FIGURE 2 | *Aloe* polysaccharides (APS) inhibits influenza virus replication in MDCK cells. **(A)** The MDCK cells were cultured with medium containing different concentrations of APS and then infected with PR8 virus [(MOI) = 1]. The virus titers of the culture supernatant collected each 12 h were determined by $TCID_{50}$. The number in parentheses represent the concentration of drug; **(B)** MDCK cells infected with PR8 virus [(MOI) = 1] were treated with 40 μg/mL of APS at different viral infection phases [before adsorption (BA), adsorption (Ad), and after adsorption (AA)] in the whole infection process. Supernatants were collected at each 24 h interval and titrated by $TCID_{50}$. **(C)** IFA detection of PR8 virus infected cells treated with APS at 4 days after inoculation. PBS treated cell wells served as the control. All values shown are presented as means ± SD from three independent experiments.

1 min. The samples were washed, dried, and examined under a JEOL-1200EX electron microscope. The isovolumetric mixture containing the virus and the STE buffer solution was used as the control.

Animal Experiment

Eighty female SPF BALB/c mice (7–8 weeks old) were randomly separated into four sterilized isolators, and each isolator contained one group (groups I to IV) comprising 20 mice. The mice were allowed to acclimatize for 3 days before the start of the experiments. All of the mice in groups I, II, and III were infected intranasally with 10^5 $TCID_{50}$ of the virus. Then, the mice in groups I and II were orally administered with 20 and 40 mg of APS per day, respectively, and the mice in group III were orally treated with isometric PBS. All mice in these three groups were administered continuously for 5 days. The mice in group IV (virus and polysaccharide free) served as the mock control. Groups I, II, III, and IV were called AP (20 mg/day), AP (40 mg/day), PBS, and Mock, respectively. The clinical symptoms and body weight loss of the mice were recorded daily up to 14 days post-infection (dpi). Additionally, another 20 normal mice in each group were placed in new isolators and then infected intranasally with 10^6 $TCID_{50}$ of the virus (a lethal dose), and the survival rate was monitored up to 7 dpi.

Detection of Pathological Changes and Viral Loads of the Lungs

Three lung tissues from each group were collected randomly at 3, 5, and 7 dpi and finely ground with a stroke-physiological saline solution at 1:10 (weight:volume). The viral titers in the lung tissues were determined by $TCID_{50}$. At 7 dpi, the pulmonary pathological section was obtained and observed through H&E staining. At the same time, viral localization in lung tissues was performed to observe the distribution of virus-infected cells in the tissue sections through immunofluorescence histochemical staining by using mouse anti-influenza NP monoclonal antibody (Abcam, Cambridge, United Kingdom).

Statistical Analysis

The data were expressed as mean ± SD, and SPSS 17.0 software was used for statistical evaluation. Duncan's multiple-range test was used to determine the differences among the groups. Statistical significance was considered at $P < 0.05$.

RESULTS

Molecular Weight (Mw) and Monosaccharide Composition of APS

Prior to the experiment, we detected the ingredient composition of APS. The Mw and monosaccharide composition of APS were determined by GPC and HPLC methods, respectively. As shown in **Tables 1**, **2**, the Number-average Mw (Mn), Weight-average Mw (Mw), and Z-average Mw (Mz) were 5354, 9198, and 16817 KDa, respectively; the monosaccharide contents of mannose, ribose, rhamnose, glucuronic acid, galacturonic acid, glucose, galactose, xylose, arabinose, and fucose in APS were 5.85, 1.50, 2.20, 5.78, 7.84, 7.57, 52.88, 0, 10.07, and 6.30%, respectively. Usually an antiviral activity of polysaccharide fractions correlates with the molecular mass of the chain (Ghosh et al., 2009). The higher the average Mw (usually ranging from 1 to 500 kDa), the higher is the antiviral activity in many cases, while above 100 kDa, no further increase in activity was observed (Witvrouw and De Clercq, 1997). From the Mw distribution result, approximately 91% of the APS fractions have a molecular weight less than 500 kDa, implying its potential antiviral effects.

APS Inhibits Replication of PR8 Virus *in vitro*

We first determined the viral replication kinetics in the MDCK cells treated with different APS concentrations to examine the antiviral characteristics of APS against the H1N1 (PR8) virus. Prior to the experiment, the toxicity of APS to MDCK cells was examined through MTT method. The results showed that APS at concentrations ranging from 20 to 640 μg/mL elicited no significant cytotoxicity on cellular activities. Conversely, the existence of APS promoted the cell growth in a dose-dependent manner (**Figure 1**). In the viral replication kinetics assay, the viral titers in the cell cultures containing 40, 80, and 160 μg/mL of APS were significantly lower than those in the control group from 24 to 72 h post-infection ($P < 0.05$; **Figure 2A**). We observed that the antiviral activity of APS remained in a typical dose-dependent manner.

We chose a minimum effective concentration above (40 μg/mL) to determine the reaction phase of APS in the viral

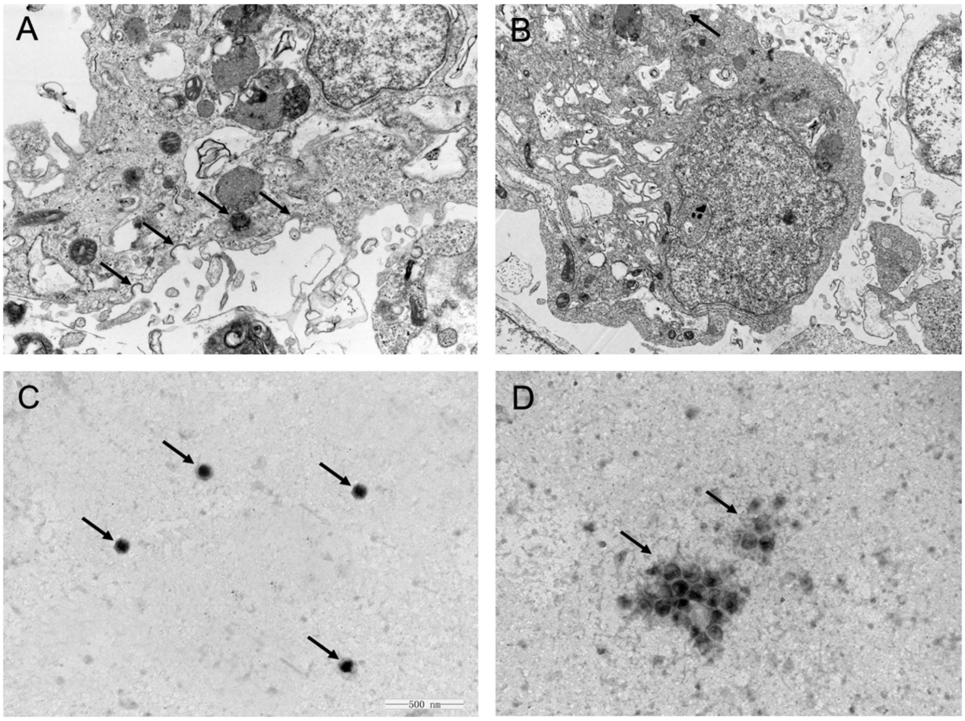


FIGURE 3 | *Aloe* polysaccharides interferes with viral adsorption to MDCK cells. MDCK monolayers were infected with MOI = 1 of PR8 virus in the absence **(A)** or presence **(B)** of APS treatment (40 μg/mL). At 4 h post-infection, the cells were centrifuged and fixed to prepare ultrathin sections, and the virions in cells were imaged via TEM (30000×). The virions on the cell membrane surface and in the endosomes were denoted by black arrowheads. Te purified PR8 virus **(C)** and the mixture of virus and APS **(D)** were imaged via TEM after negative staining (60000×). The virions were denoted by black arrowheads.

life cycle and to evaluate its antiviral effects at three artificially divided viral infection phases, namely, before adsorption (BA), adsorption (Ad), and after adsorption (AA). We determined that the viral titers of BA, Ad, and AA groups were lower than those of the non-polysaccharide-treated control group in the entire monitoring period. Minimum viral titers were detected in the Ad group ($P < 0.05$; **Figure 2B**), indicating the optimal antiviral effect of APS administered in the viral adsorption phase. The viral titers in the BA group also showed an obvious suppressive effect compared with that of the control group, indicating that the preincubation of the virus and polysaccharide could interfere with viral replication.

Similarly, the intensity and density of antigen-reactive immunofluorescence in the Ad and BA groups were significantly lower than those in the AA and control groups at 4 dpi ($P < 0.05$; **Figure 2C**). These results showed that APS could remarkably inhibit replication of PR8 virus *in vitro*, and the inhibitory effect was optimal at the viral adsorption phase.

Interaction Between APS and PR8 Virus

One feasible way for the reported plant polysaccharides to inhibit viral infection is to interfere with viral adsorption (Harden et al., 2009; Yu et al., 2017). Considering the previously presented results, we determined whether APS inhibited H1N1 infection in this manner. The viral adsorption capability of the MDCK cells treated with APS was assessed through TEM tomography. **Figure 3A** shows that several virus particles in the membrane fusion phase were observed on the cell membrane surface of the infected cells without APS treatment. By contrast, few H1N1 virus particles were observed on the membrane surface when treated with APS (**Figure 3B**). Moreover, some visible virions in the virus-infected cells were distributed densely in the endosomes (**Figure 3A**), whereas few virions were observed in the endosomes of the APS-treated cells (**Figure 3B**). This phenomenon indicated that APS reduced viral adsorption to and infection of the host cells.

Considering the previous findings (Song et al., 2013; Yang et al., 2015), we speculated that APS might act directly on the virus. As such, we conducted TEM to investigate the interaction between APS and virus particles. The scattered virus particles in the purified virus sample appeared as spheroidal structures with an inner, centrally located electron-dense core, and the average size of the virus particles was approximately 110 nm (**Figure 3C**). By contrast, the virus particles treated with APS accumulated into clusters, and the virions exhibited irregular shapes and comprised low or no electron density cores and scattered fragments around the cluster (**Figure 3D**). These results revealed a direct effect of APS on the PR8 virus particles.

APS Reduces the Pathogenicity of the PR8 Virus

We administered APS after the mice were challenged with the PR8 virus to investigate whether APS had an inhibitory activity against PR8 virus infection *in vivo*. The clinical signs and pathological changes were monitored. The observation of the clinical symptoms showed that the mice infected with PR8 virus (10^5 $TCID_{50}$) were listless and had messy, lackluster, and bristled fur (**Figure 4C**). They also exhibited mental depression and poor appetite by knocking on the cage and checking the amount of food left. However, APS administration significantly ameliorated the symptoms of the infected mice. After the treatment with 20 mg/day APS, the infected mice showed slightly messy fur and mild depression (**Figure 4A**). By contrast, the mice treated with 40 mg/day APS exhibited smooth and lustrous fur and normal behavior (**Figure 4B**), which was similar to that of the control group (**Figure 4D**). Consistent with the clinical symptoms, the body weight of the infected mice was significantly low. The loss of body weight of the mice administered with APS was significantly alleviated, and the effect of 40 mg/day APS was better than that of 20 mg/day APS ($P < 0.05$; **Figure 4E**).

Most animals and even humans who died of influenza virus infection developed acute respiratory distress syndrome, which is the most severe form of acute lung injury (Tumpey et al., 2005), and the pathogenicity is characterized by inflammatory cell accumulation, edema formation, and marked cytokine increase.

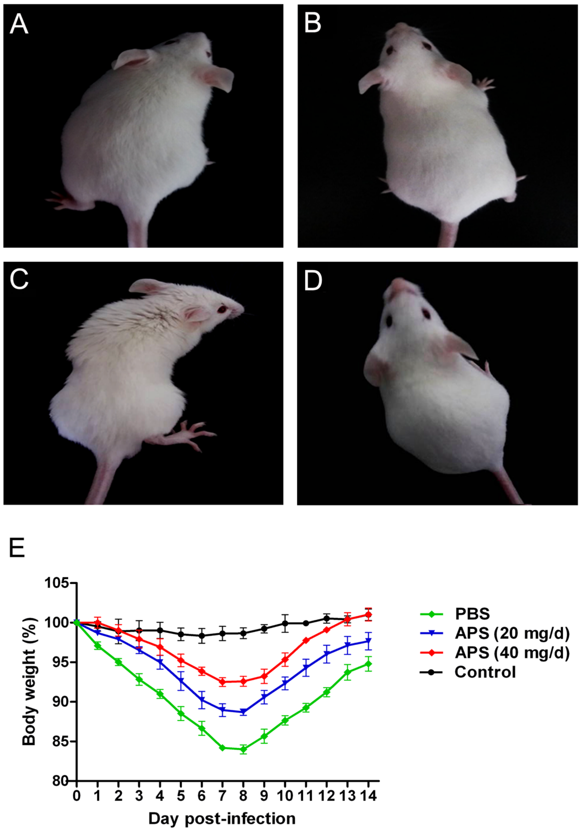


FIGURE 4 | *Aloe* polysaccharides reduces the clinical symptoms of infected mice. All 60 mice in groups I **(A)**, II **(B)**, and III **(C)** were infected intranasally with the 10^5 $TCID_{50}$ of virus. Then the mice in groups I and II were orally administrated with 20 and 40 mg of APS per day, respectively, and group III were orally administrated isometric PBS. Another 20 virus-free mice in groups IV **(D)** served as the normal control. The typical clinical symptoms were recorded at 7 dpi **(A–D)**, and the body weight loss **(E)** of animals were recorded every day up to 14 dpi.

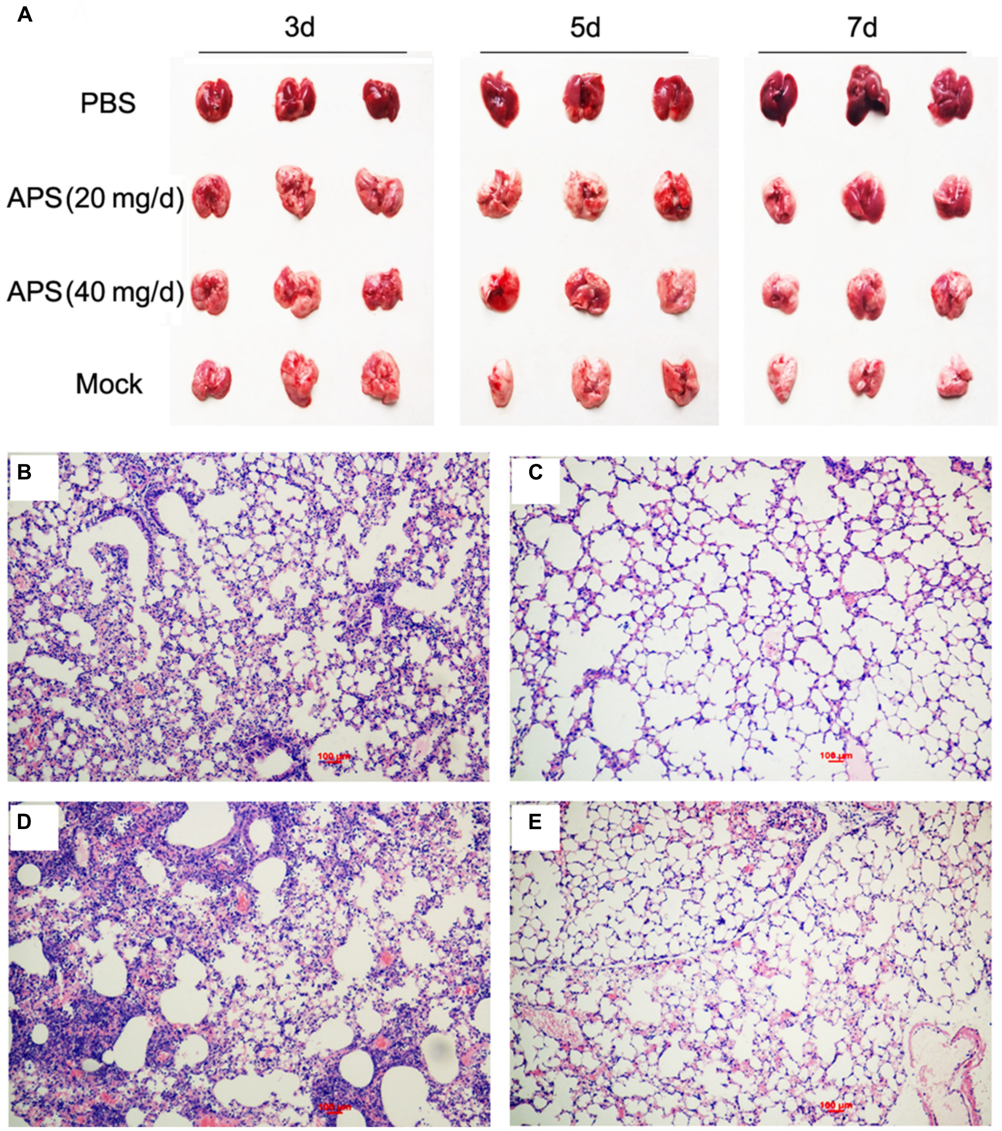


FIGURE 5 | *Aloe* polysaccharides reduces the pathological change of infected lungs. All mice in groups I, II, and III were infected intranasally with the 10^5 $TCID_{50}$ of virus. Then the mice in groups I and II were orally administrated with 20 and 40 mg of APS per day, respectively, and group III were orally administrated isometric PBS. The virus-free mice in groups IV served as the normal control. At 7 dpi, the lungs in each group were collected, and the pulmonary lesions **(A)** and pathological changes (**B**, group I; **C,** group II; **D**, group III; **E**, group IV) were compared through subjective observation and H&E staining, respectively. Scale bar, 100 μm.

A pathologic examination was performed postmortem to examine the pathological changes in the mice treated or untreated with APS against avian influenza infection. PR8-virus-infected lungs had severe lesions with extensive consolidation in all of the lobes (**Figure 5A**). The APS-treated mice did not significantly develop acute lung injury after they were infected with virus (**Figure 5A**). Furthermore, the infected lungs from different groups were excised for histopathological evaluation at 7 dpi. The microscopic lesions of the virus-infected lungs were severe peribronchiolitis and bronchopneumonia, which were characterized by edema and diffuse infiltration of inflammatory cells in the alveolar lumen and bronchioles (**Figure 5D**). However, APS treatment significantly alleviated the pulmonary histopathologic symptoms (**Figures 5B,C**). The mice treated with 40 mg/day APS had mild lesions in the lungs, which even resembled the lungs of the control group (**Figure 5E**). These results indicated that APS largely alleviated the clinical symptoms and pulmonary lesions induced by influenza virus infection.

APS Reduces the Viral Loads and Mortality of Virus-Infected Mice

We determined the degree of virus propagation in PR8-infected mice with or without APS treatment to investigate the practical virustatic effects of APS *in vivo* further. The influenza virus mainly exists in the lungs of infected animals and mediates the damage of the airway, alveolar epithelium, and alveolar

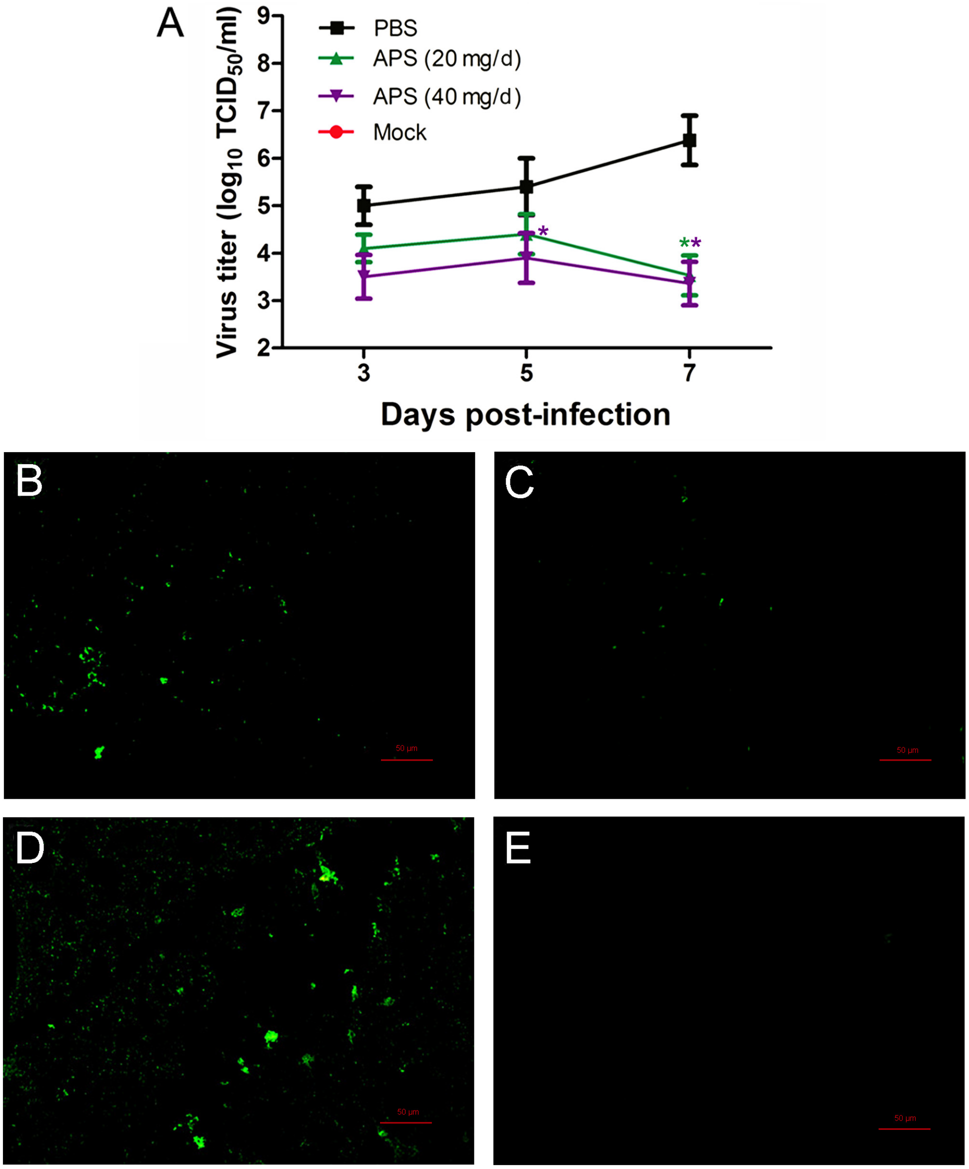


FIGURE 6 | *Aloe* polysaccharides reduces viral loads in virus-infected mice. All mice in groups I, II, and III were infected intranasally with the 10^5 $TCID_{50}$ of virus. Then the mice in groups I and II were orally administrated with 20 and 40 mg of APS per day, respectively, and group III were orally administrated isometric PBS. The virus-free mice in groups IV served as the normal control. The viral titers **(A)** in lung tissues were determined by TCID50. The viral localization in lung tissues sections was also performed through immunofluorescence histochemical staining using the mouse anti-influenza NP monoclonal antibody (200×; **B**, group I; **C**, group II; **D**, group III; **E**, group IV). Scale bar, 50 µm. An asterisk indicates that the value of the corresponding group was significantly different from that of the group III ($P < 0.05$).

endothelium (Herold et al., 2015). Thus, lung tissues from different groups were collected, and the viral titers were analyzed at 3, 5, and 7 dpi. The viral titers in the lung tissues of the 40 mg/day APS treated groups significantly decreased from 5 to 7 dpi compared with that in the control group, and those of the 20 mg/day APS treated groups by 7 dpi ($P < 0.05$; **Figure 6A**). The viral colonization in the lung tissues fixed at 7 dpi was also visualized through immunofluorescence histochemical staining by using an anti-influenza NP monoclonal antibody and fluorescent secondary antibodies. We observed intense viral antigen staining in the section of the lung tissue infected with PR8 virus (**Figure 6D**). The staining densities in the lung tissue sections of the APS-treated groups (**Figures 6B,C**) were obviously lower than those of the PBS group (**Figure 6D**), and few positively stained lung cells were observed in the section administered with 40 mg/day APS (**Figure 6C**). No viral signals were detected in the virus-free control group (**Figure 6E**). Moreover, the mortality of the mice treated with 40 mg/day APS declined by 70% compared with the mice without APS treatment after infection with a lethal dose of 10^6 $TCID_{50}$ (**Figure 7**). These results indicated that mice orally administrated with APS after virus infection could be protected from the lethal infection of the PR8 virus.

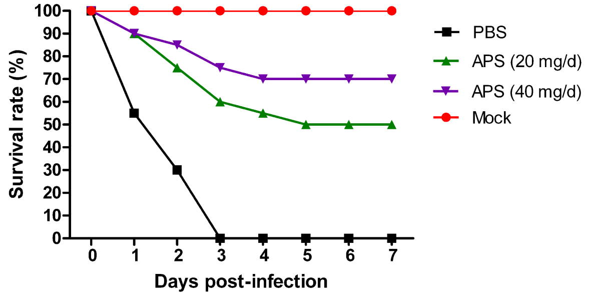


FIGURE 7 | *Aloe* polysaccharides reduces mortality in virus-infected mice. Twenty mice in groups I, II, and III were infected intranasally with the 10^6 $TCID_{50}$ of virus. Then the mice in groups I and II were orally administrated with 20 and 40 mg of APS per day, respectively, and group III were orally administrated isometric PBS. The virus-free mice in groups IV served as the normal control. The survival rate was monitored up to 7 dpi.

DISCUSSION

Seasonal influenza is a long-term threat to human health, and it causes significant morbidity and mortality every year. Exploring new ways to prevent and treat IAV infection is crucial to control influenza outbreak. In the current study, we examined the ingredient composition of APS and determined that APS exerted a significant antiviral activity to an H1N1 subtype influenza virus *in vitro* and *in vivo*. The TEM assay demonstrated that APS interacted with influenza virions, thereby preventing the attachment of the virus to the host cells. In the *in vivo* experiment, APS obviously reduced viral shedding and viral loads in mouse lungs and ameliorated the clinical symptoms and mortality of influenza virus-infected mice.

Influenza A virus is an ineradicable contagious disease causing seasonal epidemic and sporadic pandemic outbreaks that pose significant morbidity and mortality for humans and animals. IAV can be transmitted through aerosols or respiratory droplets, resulting in respiratory infections transmitted via air droplets and further inducing pneumonia (La Gruta et al., 2007; Bouvier and Palese, 2008; Kuiken et al., 2012). Herein, we determined that APS significantly induced an anti-influenza virus effect *in vitro* and an evident therapeutic effect on virus-infected mice. A dose-dependent antiviral effect was observed in cell assay. In particular, the effects were most pronounced in the early stages of infection. Given this characteristic, nasal administration may also be an effective method. Besides, APS administration alleviated pulmonary congestion, edema, and inflammatory cell infiltration and decreased the viral titers in lung tissues. These phenomena verify our speculations about the anti-influenza activity of APS. Currently, existing anti-flu drugs, such as oseltamivir and zanamivir, usually work on a particular protein of the influenza virus and can therefore easily induce drug resistance (Hata et al., 2008; Jefferson et al., 2014). However, this drawback can be effectively avoided by plant polysaccharides, which are complex macromolecules.

Acemannan and polymannose extracted from *Aloe* plants have antiviral activities as described in the Section "Introduction," and some possible mechanisms have been investigated. Saoo et al. (1996) reported the antiviral activity of *A. barbadensis* Miller against human cytomegalovirus (HCMV) and showed that the major mechanism of *Aloe* extracts in inhibiting HCMV infection involves the interference of DNA synthesis. The antiviral activity of *Aloe* emodin against IAV has also been reported (Li et al., 2014). Recently, a study shows that emodin can inhibit IAV replication and influenza viral pneumonia by activating Nrf2 signaling and inhibiting IAV-induced activation of TLR4, p38/JNK MAPK, and NF-κB pathways (Dai et al., 2017). However, the antiviral mechanism of APS is poorly understood.

To our knowledge, the antiviral mechanism of plant-origin polysaccharides is a complex process. Studies have shown three possible antiviral mechanisms of plant polysaccharides. First, polysaccharides directly interfere with viral infection in host cells (Ghosh et al., 2009; Yu et al., 2017). Second, polysaccharides induce the expression of relevant host antiviral proteins, i.e., intracellular signaling pathways (Rechter et al., 2006; Ghosh et al., 2009). Third, polysaccharides have immunoregulatory activity, for instance, the ability of APS to improve host immunity has been explored, and Lee et al. (2001) proposed that acemannan, a major carbohydrate fraction of *A. vera* gel, can promote the differentiation of immature DCs and exert immunomodulatory activity (Sun et al., 2011). Here, our *in vitro* experiments confirmed that the key time of APS action is the viral adsorption period because most virions have not yet entered host cells at this time. Therefore, APS may interfere with influenza virus adsorption by an unknown mechanism. On the basis of this assumption, we further determined the direct antiviral mechanism of APS by investigating the interaction between APS and influenza virions via the TEM assay. The results showed that APS directly affected the morphological characteristics and distribution of influenza particles. Notably, we observed a peculiar phenomenon in which the IAV particles clustered and exhibited irregular shapes and scattered fragments after these particles interacted with APS. This phenomenon was similar to that observed in our previous study of the interaction between Taishan *P. massoniana* pollen polysaccharide and avian leukemia virus (Yu et al., 2017). Therefore, the direct interaction between APSs and IAVs is a likely mechanism of inhibiting viral infection.

CONCLUSION

Polysaccharides extracted from *A. vera*, a common medicinal plant, elicited significant anti-influenza virus effects *in vivo* and *in vitro*. APS could directly interact with PR8 (H1N1) influenza virus particles to prevent virus particle adsorption. Our previous studies have also shown that APS can increase the activity of the immune system, which is another important mechanism of antiviral infection. In future studies, more virus strains and hosts should be used for further validation, and signaling pathways associated with APS interaction will be investigated to explain their unknown anti-influenza mechanism. This study provided a critical theoretical basis for the development of APS as a novel anti-influenza drug.

AUTHOR CONTRIBUTIONS

KW, JZ, and ZS designed the research. ZS, CY, WW, GY, TZ, and LZ performed the research. ZS, CY, KW, and JZ analyzed the data and wrote the paper.

FUNDING

This project was funded by the Natural Science Foundation of Shandong Province (ZR2017BH101 and ZR2014JL051), the National Natural Science Foundation of China (31702254), Administration of Traditional Chinese Medicine of Shandong Province (2015-258), and High-level Project Cultivation Plan of Taishan Medical University (2015GCC02).

REFERENCES

Bouvier, N. M., and Palese, P. (2008). The biology of influenza viruses. *Vaccine* 26, D49–D53.

Dai, J.-P., Wang, Q.-W., Su, Y., Gu, L.-M., Zhao, Y., Chen, X.-X., et al. (2017). Emodin Inhibition of Influenza A Virus Replication and Influenza Viral Pneumonia via the Nrf2, TLR4, p38/JNK and NF-kappaB Pathways. *Molecules* 22:E1754. doi: 10.3390/molecules22101754

Gauntt, C., Wood, H., McDaniel, H., and McAnalley, B. (2000). Aloe polymannose enhances anti-coxsackievirus antibody titres in mice. *Phytother. Res.* 14, 261–266. doi: 10.1002/1099-1573(200006)14:4<261::AID-PTR579>3.0.CO;2-A

Ghosh, T., Chattopadhyay, K., Marschall, M., Karmakar, P., Mandal, P., and Ray, B. (2009). Focus on antivirally active sulfated polysaccharides: from structure–activity analysis to clinical evaluation. *Glycobiology* 19, 2–15. doi: 10.1093/glycob/cwn092

Harden, E. A., Falshaw, R., Carnachan, S. M., Kern, E. R., and Prichard, M. N. (2009). Virucidal activity of polysaccharide extracts from four algal species against herpes simplex virus. *Antiviral Res.* 83, 282–289. doi: 10.1016/j.antiviral.2009.06.007

Hata, K., Koseki, K., Yamaguchi, K., Moriya, S., Suzuki, Y., Yingsakmongkon, S., et al. (2008). Limited inhibitory effects of oseltamivir and zanamivir on human sialidases. *Antimicrob. Agents Chemother.* 52, 3484–3491. doi: 10.1128/AAC.00344-08

Herold, S., Becker, C., Ridge, K. M., and Budinger, G. S. (2015). Influenza virus-induced lung injury: pathogenesis and implications for treatment. *Eur. Respir. J.* 45, 1463–1478. doi: 10.1183/09031936.00186214

Hibino, A., Kondo, H., Masaki, H., Tanabe, Y., Sato, I., Takemae, N., et al. (2017). Community-and hospital-acquired infections with oseltamivir-and peramivir-resistant influenza A (H1N1) pdm09 viruses during the 2015–2016 season in Japan. *Virus Genes* 53, 89–94. doi: 10.1007/s11262-016-1396-9

Iljazovic, E., Zulcic-Nakic, V., Latifagic, A., Sahimpasic, A., Omeragic, F., and Avdic, S. (2006). 245 ORAL Efficacy in treatment of cervical HRHPV infection by combination of interferon, Aloe vera and propolis gel associated with different cervical lesion. *Eur. J. Surg. Oncol.* 32, S73–S139. doi: 10.1016/S0748-7983(06)70680-1

Jefferson, T., Jones, M. A., Doshi, P., Del Mar, C. B., Hama, R., Thompson, M., et al. (2014). Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. *Sao Paulo Med. J.* 132, 256–257. doi: 10.1590/1516-3180.20141324T2

Kahlon, J., Kemp, M., Carpenter, R., McAnalley, B., McDaniel, H., and Shannon, W. (1991a). Inhibition of AIDS virus replication by acemannan in vitro. *Mol. Biother.* 3, 127–135.

Kahlon, J., Kemp, M., Yawei, N., Carpenter, R., Shannon, W., and McAnalley, B. (1991b). In vitro evaluation of the synergistic antiviral effects of acemannan in combination with azidothymidine and acyclovir. *Mol. Biother.* 3, 214–223.

Kuiken, T., Riteau, B., Fouchier, R., and Rimmelzwaan, G. (2012). Pathogenesis of influenza virus infections: the good, the bad and the ugly. *Curr. Opin. Virol.* 2, 276–286. doi: 10.1016/j.coviro.2012.02.013

Kumar, S., and Tiku, A. B. (2016). Immunomodulatory potential of acemannan (polysaccharide from Aloe vera) against radiation induced mortality in Swiss albino mice. *Food Agric. Immunol.* 27, 72–86. doi: 10.1080/09540105.2015.1079594

La Gruta, N. L., Kedzierska, K., Stambas, J., and Doherty, P. C. (2007). A question of self-preservation: immunopathology in influenza virus infection. *Immunol. Cell Biol.* 85, 85–92. doi: 10.1038/sj.icb.7100026

Lambert, L. C., and Fauci, A. S. (2010). Influenza vaccines for the future. *N. Engl. J. Med.* 363, 2036–2044. doi: 10.1056/NEJMra1002842

Langmead, L., Makins, R., and Rampton, D. (2004). Anti-inflammatory effects of aloe vera gel in human colorectal mucosa in vitro. *Aliment. Pharmacol. Ther.* 19, 521–527. doi: 10.1111/j.1365-2036.2004.01874.x

Lee, J.-B., Tanikawa, T., Hayashi, K., Asagi, M., Kasahara, Y., and Hayashi, T. (2015). Characterization and biological effects of two polysaccharides isolated from *Acanthopanax sciadophylloides*. *Carbohydr. Polym.* 116, 159–166. doi: 10.1016/j.carbpol.2014.04.013

Lee, J. K., Lee, M. K., Yun, Y.-P., Kim, Y., Kim, J. S., Kim, Y. S., et al. (2001). Acemannan purified from Aloe vera induces phenotypic and functional maturation of immature dendritic cells. *Int. Immunopharmacol.* 1, 1275–1284. doi: 10.1016/S1567-5769(01)00052-2

Leung, K., Lipsitch, M., Yuen, K. Y., and Wu, J. T. (2017). Monitoring the fitness of antiviral-resistant influenza strains during an epidemic: a mathematical modelling study. *Lancet Infect. Dis.* 17, 339–347. doi: 10.1016/S1473-3099(16)30465-0

Li, S.-W., Yang, T.-C., Lai, C.-C., Huang, S.-H., Liao, J.-M., Wan, L., et al. (2014). Antiviral activity of aloe-emodin against influenza A virus via galectin-3 up-regulation. *Eur. J. Pharmacol.* 738, 125–132. doi: 10.1016/j.ejphar.2014.05.028

Liu, S. S., Jiao, X. Y., Wang, S., Su, W. Z., Jiang, L. Z., Zhang, X., et al. (2017). Susceptibility of influenza A (H1N1)/pdm2009, seasonal A (H3N2) and B viruses to Oseltamivir in Guangdong, China between 2009 and 2014. *Sci. Rep.* 7:8488. doi: 10.1038/s41598-017-08282-6

Ly, S., Horwood, P., Chan, M., Rith, S., Sorn, S., Oeung, K., et al. (2017). Seroprevalence and transmission of human influenza A (H5N1) virus before and after virus reassortment, Cambodia, 2006–2014. *Emerg. Infect. Dis.* 23:300. doi: 10.3201/eid2302.161232

Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J. Immunol. Methods* 65, 55–63. doi: 10.1016/0022-1759(83)90303-4

Nguyen, T. L., Chen, J., Hu, Y., Wang, D., Fan, Y., Wang, J., et al. (2012). In vitro antiviral activity of sulfated *Auricularia auricula* polysaccharides. *Carbohydr. Polym.* 90, 1254–1258. doi: 10.1016/j.carbpol.2012.06.060

Nichol, K. L., and Treanor, J. J. (2006). Vaccines for seasonal and pandemic influenza. *J. Infect. Dis.* 194(Suppl. 2), S111–S118. doi: 10.1086/507544

Pan, M., Gao, R., Lv, Q., Huang, S., Zhou, Z., Yang, L., et al. (2016). Human infection with a novel, highly pathogenic avian influenza A (H5N6) virus: virological and clinical findings. *J. Infect.* 72, 52–59. doi: 10.1016/j.jinf.2015.06.009

Queiroz, K., Medeiros, V., Queiroz, L., Abreu, L., Rocha, H., Ferreira, C., et al. (2008). Inhibition of reverse transcriptase activity of HIV by polysaccharides of brown algae. *Biomed. Pharmacother.* 62, 303–307. doi: 10.1016/j.biopha.2008.03.006

Radha, M. H., and Laxmipriya, N. P. (2015). Evaluation of biological properties and clinical effectiveness of Aloe vera: a systematic review. *J. Tradit. Complement. Med.* 5, 21–26. doi: 10.1016/j.jtcme.2014.10.006

Rechter, S., König, T., Auerochs, S., Thulke, S., Walter, H., Dörnenburg, H., et al. (2006). Antiviral activity of Arthrospira-derived spirulan-like substances. *Antiviral Res.* 72, 197–206. doi: 10.1016/j.antiviral.2006.06.004

Saoo, K., Miki, H., Ohmori, M., and Winters, W. (1996). Antiviral activity of aloe extracts against cytomegalovirus. *Phytother. Res.* 10, 348–350. doi: 10.1002/(SICI)1099-1573(199606)10:4<348::AID-PTR836>3.0.CO;2-2

Song, X., Yin, Z., Li, L., Cheng, A., Jia, R., Xu, J., et al. (2013). Antiviral activity of sulfated *Chuanminshen violaceum* polysaccharide against duck enteritis virus in vitro. *Antiviral Res.* 98, 344–351. doi: 10.1016/j.antiviral.2013.03.012

Sun, Z., Wei, K., Yan, Z., Zhu, X., Wang, X., Wang, H., et al. (2011). Effect of immunological enhancement of aloe polysaccharide on chickens immunized with *Bordetella avium* inactivated vaccine. *Carbohydr. Polym.* 86, 684–690. doi: 10.1016/j.carbpol.2011.05.012

Swayne, D. E. (2009). Avian influenza vaccines and therapies for poultry. *Comp. Immunol. Microbiol. Infect. Dis.* 32, 351–363. doi: 10.1016/j.cimid.2008.01.006

Syed, T. A., Cheema, K. M., Ahmad, S. A., and Holt, A. H. Jr. (1996). Aloe vera extract 0.5% in hydrophilic cream versus Aloe vera gel for the management of genital herpes in males. A placebo-controlled, double-blind, comparative study. *J. Eur. Acad. Dermatol. Venereol.* 7, 294–295.

Taubenberger, J. K., and Morens, D. M. (2017). H5Nx panzootic Bird Flu—influenza's newest worldwide evolutionary tour. *Emerg. Infect. Dis.* 23:340. doi: 10.3201/eid2302.161963

Tumpey, T. M., Basler, C. F., Aguilar, P. V., Zeng, H., Solórzano, A., Swayne, D. E., et al. (2005). Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* 310, 77–80. doi: 10.1126/science.1119392

Witvrouw, M., and De Clercq, E. (1997). Sulfated polysaccharides extracted from sea algae as potential antiviral drugs. *Gen. Pharmacol.* 29, 497–511. doi: 10.1016/S0306-3623(96)00563-0

Yagi, A., and Byung, P. Y. (2015). Immune modulation of Aloe vera: acemannan and gut microbiota modulator. *J. Gastroenterol. Hepatol. Res.* 4, 1707–1721. doi: 10.17554/j.issn.2224-3992.2015.04.525

Yang, S., Wei, K., Jia, F., Zhao, X., Cui, G., Guo, F., et al. (2015). Characterization and biological activity of Taishan *Pinus* massoniana pollen polysaccharide *in vitro*. *PLoS One* 10:e0115638. doi: 10.1371/journal.pone.0115638

Ylipalosaari, P., Ala-Kokko, T. I., Laurila, J., Ahvenjärvi, L., and Syrjälä, H. (2017). ICU-treated influenza A (H1N1) pdm09 infections more severe post pandemic than during 2009 pandemic: a retrospective analysis. *BMC Infect. Dis.* 17:728. doi: 10.1186/s12879-017-2829-3

Yu, C., Wei, K., Liu, L., Yang, S., Hu, L., Zhao, P., et al. (2017). Taishan *Pinus* massoniana pollen polysaccharide inhibits subgroup J avian leucosis virus infection by directly blocking virus infection and improving immunity. *Sci. Rep.* 7:44353. doi: 10.1038/srep44353

Zandi, K., Zadeh, M. A., Sartavi, K., and Rastian, Z. (2007). Antiviral activity of *Aloe* vera against herpes simplex virus type 2: an *in vitro* study. *Afr. J. Biotechnol.* 6.

Conflict of Interest Statement: 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.

frontiers in Microbiology

PERSPECTIVE
published: 26 October 2018
doi: 10.3389/fmicb.2018.02604

Manifold Routes to a Nucleus

Heather L. Hendrickson[1]* *and Anthony M. Poole*[2,3,4]*

[1] Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand, [2] Bioinformatics Institute, The University of Auckland, Auckland, New Zealand, [3] Te Ao Mârama/Centre for Fundamental Inquiry, The University of Auckland, Auckland, New Zealand, [4] School of Biological Sciences, The University of Auckland, Auckland, New Zealand

It is widely assumed that there is a clear distinction between eukaryotes, with cell nuclei, and prokaryotes, which lack nuclei. This suggests the evolution of nuclear compartmentation is a singular event. However, emerging knowledge of the diversity of bacterial internal cell structures suggests the picture may not be as black-and-white as previously thought. For instance, some members of the bacterial PVC superphylum appear to have nucleus-like compartmentation, where transcription and translation are physically separated, and some jumbophages have recently been shown to create nucleus-like structures within their Pseudomonad hosts. Moreover, there is also tantalizing metagenomic identification of new Archaea that carry homologs of genes associated with internal cell membrane structure in eukaryotes. All these cases invite comparison with eukaryote cell biology. While the bacterial cases of genetic compartmentation are likely convergent, and thus viewed by many as not germane to the question of eukaryote origins, we argue here that, in addressing the broader question of the evolution of compartmentation, other instances are at least as important: they provide us with a point of comparison which is critical for a more general understanding of both the conditions favoring the emergence of intracellular compartmentation of DNA and the evolutionary consequences of such cellular architecture. Finally, we consider three classes of explanation for the emergence of compartmentation: physical protection, crosstalk avoidance and nonadaptive origins.

Keywords: nucleus, compartmentation, planctomycetes, jumbophage, Asgard, RNA avoidance

OPEN ACCESS

Edited by:
Arshan Nasir,
COMSATS Institute of Information Technology, Pakistan

Reviewed by:
Patrick Forterre,
Institut Pasteur, France
Masaharu Takemura,
Tokyo University of Science, Japan

****Correspondence:***
Heather L. Hendrickson
h.hendrickson@massey.ac.nz
Anthony M. Poole
a.poole@auckland.ac.nz

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

Received: *30 July 2018*
Accepted: *11 October 2018*
Published: *26 October 2018*

Citation:
Hendrickson HL and Poole AM (2018) Manifold Routes to a Nucleus. Front. Microbiol. 9:2604. doi: 10.3389/fmicb.2018.02604

THE CONUNDRUM OF THE EUKARYOTE NUCLEUS

The eukaryote nucleus (**Figure 1A**) is one of the most remarkable structures in biology. It is home to the major part of the genetic material in eukaryotic cells, and is conserved across all eukaryotes, which share a core set of genes for the nuclear pore complex (Mans et al., 2004; Bapteste et al., 2005; Neumann et al., 2010) and nucleocytoplasmic transport (Koumandou et al., 2013). Much speculation on the origin of the eukaryote nucleus has been published, and models fall into two broad classes: endosymbiotic and autogenous (Martin et al., 2001). In endosymbiotic models, the nucleus is proposed to have evolved from a once free-living cell or from a virus (Bell, 2001; Takemura, 2001; Forterre and Prangishvili, 2009; Forterre and Gaia, 2016), whereas in most autogenous models, the nucleus evolved through internal changes that led to compartmentalization of the genetic material (Devos et al., 2004). The question of eukaryote origins has been given new fuel via the recent metagenomic identification of a new group of Archaea, called Asgard (Spang et al., 2015; Zaremba-Niedzwiedzka et al., 2017), which encode putative homologs to eukaryotic nucleocytoplasmic transport genes (Klinger et al., 2016). If the function of these archaeal genes can be established and related to cell ultrastructure, it will assuredly improve our understanding the evolution of eukaryotes (Dey et al., 2016), given a shared evolutionary history with the Archaea.

However, it is difficult to fully untangle the evolution of singular events (De Duve, 2005): did nuclear compartments evolve only once, thus requiring some special evolutionary explanation (Lane, 2011), or can they be understood with reference to known processes (Poole and Penny, 2007; Booth and Doolittle, 2015)? Researchers trying to understand the origins of life know this dilemma well: all cellular life on earth derives from a single origin. While it is possible that there were multiple origins of life on Earth (Davies and Lineweaver, 2005), there is no evidence that extant cells derive from independent origins (Penny et al., 2003). Thus, reconstructing the origin and evolution of life on earth provides us with the history of a specific instance. To understand the broader process of how life originates, it would be helpful – perhaps essential – to be able to compare life on earth with life that derives from one or more independent origins (McKay, 2004). This would help us to understand whether there is only one way that life can originate, or if there are there many routes.

In this spirit, we consider here the broader question of the origins of genetic compartmentation, of which the origin of the eukaryote nucleus has widely been assumed to represent the single instance. In contrast to the origin of life, the origin of genetic compartmentation is no longer a singularity; multiple forms of genetic compartmentation have now been observed, including among bacteria. This enables us to pose a more general question, of which the origin of the eukaryote nucleus is a specific historical instance: what forces drive the compartmentalization of genetic information within cells?

NUCLEUS-LIKE COMPARTMENTATION IS FOUND OUTSIDE OF EUKARYOTES

Members of the bacterial "PVC" superphylum, which comprises the Planctomycetes, Verrucomicrobia and Chlamydiae, appear unusual among bacteria in that a number have internal membranes (Lee et al., 2009; Fuerst and Sagulenko, 2011). The most studied member of this group, the planctomycete *Gemmata obscuriglobus*, possesses a nucleus-like compartment that contains DNA (Fuerst, 2005; Sagulenko et al., 2014), and exhibits physical separation of transcription and translation (Gottshall et al., 2014). Moreover, this nucleus-like compartment possesses membrane-spanning pores that are about a third the diameter of the eukaryote nuclear pore complex, and bear unmistakeable resemblance to eukaryote nuclear pores, both in overall architecture (**Figure 1B**), and in regard to the kinds of protein domains (beta-propellers, alpha solenoids) identified within its protein constituents (Sagulenko et al., 2017). Interestingly, it does appear that some ribosomes are found in the same compartment as the genetic material (Fuerst and Sagulenko, 2011), so the extent to which translation and transcription are separated by this compartmental barrier is unclear. While some have speculated that the nucleus-like structure in planctomycetes such as *G. obscuriglobus* could belie an ancient origin for nuclear compartmentation (Fuerst and Sagulenko, 2012; Staley and Fuerst, 2017), it seems more likely to be a stunning case of evolutionary convergence (Sagulenko et al., 2017), though it is worth noting that the exact nature of this compartmentation is a matter of ongoing debate (Acehan et al., 2014; Sagulenko et al., 2014, 2017; Boedeker et al., 2017).

Of importance here is whether the probable convergence of the intracellular compartmentation of *Gemmata* renders such compartmentation irrelevant to evolution of the eukaryote nucleus (McInerney et al., 2011). From an historical perspective, this is true: an understanding of the specific evolutionary history of bird wings does not shed light on the evolutionary history of flight in bats. But consider for a moment the value of a separate origin: comparison of the two kinds of flight reveals separate evolutionary solutions to the same problem. If planctomycete nuclear-like compartmentation does hold up to closer scrutiny, then it seems difficult to explain it via existing models for the origin of the nucleus. For instance, the suggestion that intron invasion necessitated a separation of splicing and translation (Lopez-Garcia and Moreira, 2006; Martin and Koonin, 2006) does not explain the cellular organization of *Gemmata*, not least because Planctomycetes lack an intron-exon gene organization and a mechanism for splicing. Thus, while Asgard lineages promise to shed light on the intermediate steps in eukaryogenesis (Eme et al., 2017; Fournier and Poole, 2018), the independent emergence in Planctomycetes of a nucleus-like compartment with separated transcription and translation may instead shed light on the range of mechanisms by which this type of architecture may evolve.

JUMBOPHAGE FORM PROTEINACEOUS "NUCLEUS-LIKE" COMPARTMENTS

Another fascinating case of genetic compartmentation in bacteria has recently been discovered. During infection of *Pseudomonas chlororaphis* by jumbophage 201$_{\Phi}$2-1, a genetic compartment is formed (Chaikeeratisak et al., 2017b). Jumbophages have very large genomes, the largest of which currently stands at 497.5 kb (Bacillus phage G; GenBank accession: JN638751) (Hendrix, 2009; Yuan and Gao, 2017). Following host infection, jumbophage 201$_{\Phi}$2-1 forms a dynamic protein-based compartment within the body of the bacterial cell (**Figure 1C**). Within this structure, phage DNA is replicated and transcribed whilst phage mRNA is transported outside of the compartment where it is translated, enabling the construction of phage particles (Chaikeeratisak et al., 2017b). In this regard, this protein-based partitioning is "nucleus-like"—transcription is separated from translation—though it is clearly not homologous to the eukaryote nucleus. This may be a more widespread phenomenon; phages $_{\Phi}$PA3 and $_{\Phi}$KZ, that infect *P. aeruginosa*, also generate this kind of nucleus-like structure. In all three cases, infection involves PhuZ, a tubulin homolog that forms a spindle that serves to position the nucleus, much like tubulin in

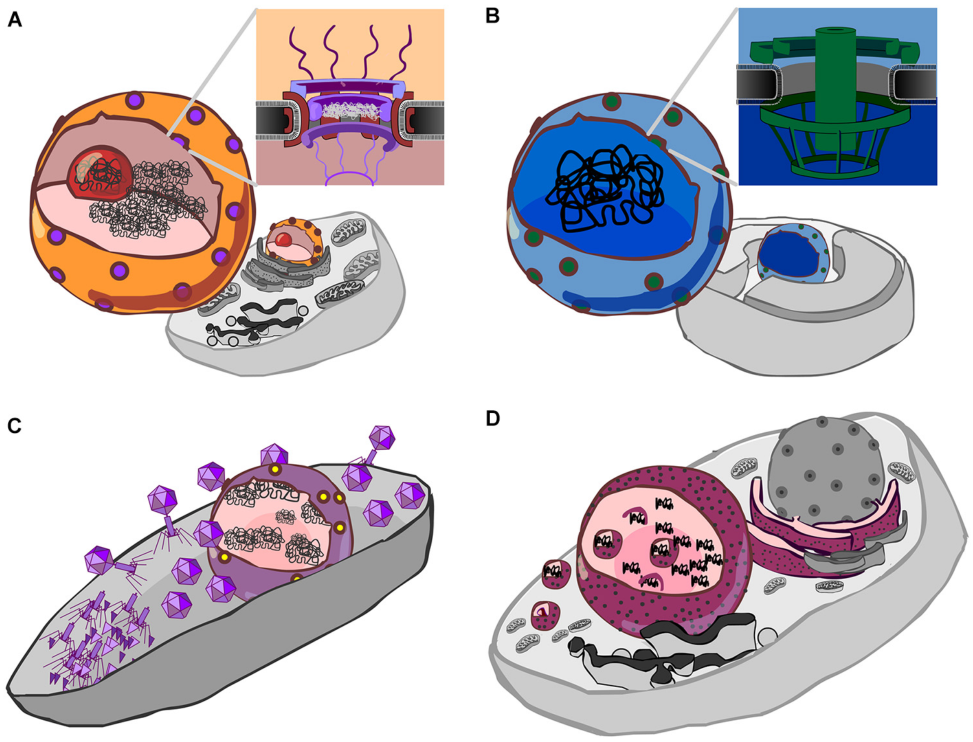


FIGURE 1 | Forms of genetic compartmentation. **(A)** Cartoon depiction of a eukaryote cell and nucleus (orange) with nuclear pores (purple), nucleolus (red) and DNA (black). Inset: a schematic cross-section of the nuclear pore complex with a central plug, cytoplasmic fibrils above, and basket structure below the plane of the membrane. The inner and outer membranes of the nuclear envelope are continuous. **(B)** Cartoon depiction of a *Gemmata obscuriglobus* cell showing internal membranes (Fuerst and Sagulenko, 2011), a proposed nucleus-like structure (blue) which contains DNA, and pores (green). Inset: a schematic of the nuclear pore-like structure reported by Sagulenko et al. (2017). **(C)** Cartoon detailing the Jumbo phage proteinaceous nucleus-like structure (purple) containing phage DNA (black), after Chaikeeratisak et al. (2017b). Protein synthesis and viral assembly occurs outside this structure, whereas DNA replication and transcription occur within. This suggests the presence of a mechanism for both export of RNA and DNA, depicted here as a single complex by yellow dots. **(D)** Cartoon detailing co-option of the endoplasmic reticulum by Mimivirus, a Nucleocytpplasmic Large DNA virus (NCLDV). Here, the rough endoplasmic reticulum (maroon with black speckles) is recruited to form an encapsulating virus factory (Mutsafi et al., 2013). A complex series of events then leads to formation of viral capsids which are then encapsulated by protein-coated membranes (Kuznetsov et al., 2013) (depicted by small magenta spheres).

eukaryotes, and gp105, a nuclear shell protein (Chaikeeratisak et al., 2017a,b). Homologs of both are detectable in 5 of the 8 complete *Pseudomonas* jumbophage in Genbank, plus in two recently sequenced jumbophages from New Zealand (HH, unpublished observations), suggesting this may be a more widespread process.

Given that the origin of the nucleus remains an unsolved problem in biology, the formation of a proteinaceous compartment during jumbophage infection invites parallels to viral-origin models for the eukaryote nucleus. In this class of model, cooption of internal membranes by infecting viruses drove stable formation of a nuclear compartment (Bell, 2001, 2009; Takemura, 2001; Forterre and Gaia, 2016). The comparison that is frequently made is with members of the NucleoCytoplasmic Large DNA Viruses (NCLDV) group, which recruit internal host membranes (Kuznetsov et al., 2013; Mutsafi et al., 2013) to create an intracellular compartment (**Figure 1D**). By contrast, the *de novo* formation of the jumbophage proteinaceous compartment indicates that compartments can be directly generated by virus infection. As this process happens in unrelated, genomically large, bacterial and eukaryote viruses, a more general question is this: might there a benefit to compartmentalization of the genetic material of large viruses? And what of compartmentalization more generally?

WHAT DRIVES THE COMPARTMENTATION OF GENETIC INFORMATION?

Indeed, the evolutionarily independent forms of compartmentation in eukaryotes, the PVC superphylum, NCLDV viral factories and jumbophage nuclei enable us to consider more generally what might drive compartmentation of genetic material. The open questions are: do any of these separate instances share a common driver or are they the result of unrelated evolutionary pressures? Is compartmentation neutral or non-adaptive? To begin to address these, we next consider three broad categories of explanation that might account for the emergence of compartmentation: physical protection, crosstalk avoidance and non-adaptive origins; we explain each in turn.

Protection

Chemical Protection

Keeping DNA physically separate from certain enzymatic reactions could act as a means of limiting chemical damage to the DNA. Under this model, damage would be prevented because the DNA is physically separated from damaging chemistries. It seems highly unlikely that the nuclear pores, which allow passive diffusion of small (30–60 kDa) molecules (Timney et al., 2016), or their smaller planctomycete analogues (Sagulenko et al., 2017) substantially reduce chemically induced DNA damage; in eukaryotes, mitochondria and chloroplasts are key sites of redox chemistry that may damage DNA, and this may be but one driver of gene relocation to the nucleus (Race et al., 1999). In the anammoxosome in planctomycetes (the site of anaerobic ammonia oxidation), specialized compartments contain the toxic intermediate of annamox chemistry (Sinninghe Damste et al., 2002) (**Figure 2A**), and this is much the same with eukaryotic peroxisomes (Gabaldon, 2010). Rather than sequestering DNA from potentially damaging chemistries, it seems the opposite occurs: it is the damaging chemistries that are compartmentalized. In lineages adapted to high levels of mutational damage, such as *Deinococcus radiodurans*, protection is via high ploidy (providing genome redundancy) and a resistance of the proteome to damage (without working repair proteins, there can be no DNA repair) (Daly et al., 2007; Slade et al., 2009; Krisko and Radman, 2013). It therefore seems that compartmentalisation of DNA is at best an infrequent solution to chemical or environmental sources of DNA damage.

Biological Protection

A more plausible argument for compartmentation of DNA may be as a means of protection from biological agents. In the case of viruses, a physical barrier may provide a powerful means by which to thwart host defenses. Many anti-phage systems have evolved and many of these rely on physical access to phage DNA in order to act, including CRISPR-Cas, restriction endonucleases and abortive infection systems (Labrie et al., 2010; Koonin et al., 2017). The discovery of the jumbophage "nucleus" prompted speculation that encapsulation of the phage DNA during infection provides an effective physical barrier for guarding the DNA from host defense systems (**Figure 2B**) (Chaikeeratisak et al., 2017b). In a recent article posted to BioR$_x$iv, Mendoza et al. (2018) directly test this, and report that jumbophage ΦKZ is resistant to several CRISPR-Cas systems and type I restriction-modification systems, but is sensitive to CRISPR-Cas13a, which can target phage mRNA that must exit the shell for translation. Clearly, this is not a general strategy among viruses, but it might be the case that large viruses represent both a bigger genomic target and that they have enough genomic space to be able to carry larger suites of genes devoted to thwarting host defenses. In addition, their burst sizes are an order of magnitude lower than many smaller viruses, so physical protection may simply be less of an issue for viruses where infection yields many more virions. One might therefore expect physical barriers to be a mechanism common among large viruses. Certainly, this appears consistent with the lifecycle of NCLDVs, which co-opt internal membrane material during infection. That said, it may be difficult to tease apart protection and organization; while the large viruses of eukaryotes often form "viral factories" by coopting host membranes, this process appears to vary considerably (Novoa et al., 2005) and it is notable that the viral factory of Mimivirus appears not to be surrounded by a membrane (Suzan-Monti et al., 2007), though, having said that, it does appear that the capsid self-assembles on vesicles that derive from the nuclear membrane or rough endoplasmic reticulum (Kuznetsov et al., 2013).

More generally, nuclear envelopes in eukaryotes and planctomycetes could reduce opportunities for integration of foreign DNA, thus acting as a possible barrier to infection or invasion of genetic elements. At the level of phylum, available data are consistent with this interpretation; Jeong et al. (2016) recently calculated HGT indices ($N_{horizontal}/N_{total}$ genes) for major bacterial and archaeal phyla, and report that Planctomycetes, Verrucomicrobia and Chlamydiae all have HGT indices well below the global median, with the first two phyla being among the lowest in their analysis. This could be further tested by examining rates of HGT in lineages with and without nuclear compartmentation; the PVC superphylum could be a good test-bed for this as there are a range of cellular architectures across this group (Fuerst and Sagulenko, 2011). Nevertheless, eukaryote horizontal gene transfer rates are appreciable (Leger et al., 2018), it is unclear that gene transfer is selected against, and, relative to prokaryotes, eukaryotic nuclear genomes do not show evidence of an improved capacity to repel invasive genetic elements; genomic evidence suggests quite the opposite in fact (Werren, 2011). Accumulation of such elements appear instead to be linked to the mode of reproduction (spread of elements is most effective under sexual outcrossing) (Arkhipova and Meselson, 2000) and the capacity to purge slightly deleterious mutations (Lynch et al., 2011). That said, the barrier provided by the nucleus does act to slow the invasion of retroelements that must enter the nucleus to replicate. Several yeast nucleoporin genes are under positive selection, and genetic changes to these impact the propensity of Ty retrotransposons to replicate in their host (Rowley et al., 2018). This indicates that the nuclear envelope can directly impact replicative spread of genetic elements in the nucleus. It is thus plausible that the nuclear envelope first evolved to slow the spread of transposable elements following the evolution of meiosis.

Crosstalk Avoidance

Crosstalk, i.e., unintended interaction between molecules, has been considered as a possible mechanism that could drive nuclear compartmentation (**Figure 2C**). One model for this is avoidance of ribosomal readthrough into introns (Lopez-Garcia and Moreira, 2006; Martin and Koonin, 2006). Under this model, crosstalk is a temporal problem, where interactions must occur in a specific order; the physical separation of transcription and translation may thus provide a temporal means for splicing to complete before translation begins, therefore preventing the formation of aberrant proteins via translational readthrough into unspliced introns, which may have toxic effects. The model is plausible, but, from a genetic perspective, it would work only under conditions of sexual reproduction, since intron invasion is

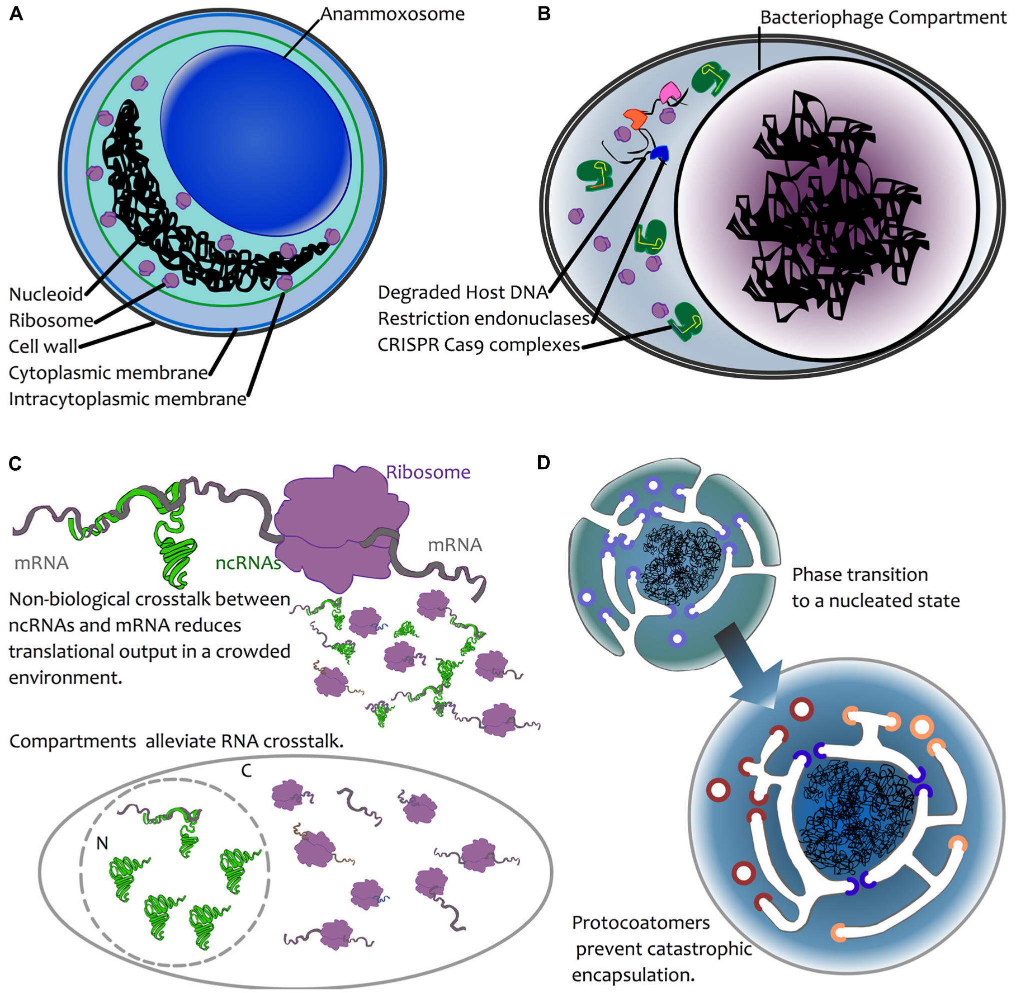

A
Anammoxosome
Nucleoid
Ribosome
Cell wall
Cytoplasmic membrane
Intracytoplasmic membrane
B
Bacteriophage Compartment
Degraded Host DNA
Restriction endonuclases
CRISPR Cas9 complexes
C
Ribosome
mRNA
ncRNAs
mRNA
Non-biological crosstalk between ncRNAs and mRNA reduces translational output in a crowded environment.
Compartments alleviate RNA crosstalk.
C
N
D
Phase transition to a nucleated state
Protocoatomers prevent catastrophic encapsulation.

FIGURE 2 | Possible mechanisms for compartmentalization of genetic material. **(A)** Chemical sequestration—as seen in annamox bacteria—where anaerobic ammonium oxidation is sequestered away from the rest of the cell in a dedicated compartment, the anammoxosome (van Niftrik, 2013). **(B)** Biological sequestration—the creation of dedicated viral compartments by large viruses (NCLDVs, Jumbophage) may prevent the action of host defenses or competition by other infectious agents by excluding these during biogenesis. For simplicity, an infected bacterial cell is depicted here, with CRISPR-Cas (green) and restriction endonuclease-based defense mechanisms excluded from the compartment. Restriction endonucleases are depicted in blue, pink orange. **(C)** Biological avoidance. Top: mRNA (grey) may interact with a ribosome (purple), leading to translation. In the case of ncRNA-mRNA interactions, these may occur through chance base-pairing, leading to reduced translational output. This is depicted by the smaller molecules, where some mRNAs interact with ribosomes, and some are trapped in unproductive interactions with ncRNAs (see text for details). In both prokaryotes and eukaryotes, avoidance of such crosstalk is a result of selection against interactions (Marco, 2015; Umu et al., 2016). Bottom: compartmentation is hypothesized to alleviate crosstalk interactions due to physical separation. The presence of ribosomes in the cytoplasm (C) and ncRNAs in the nucleus (N) reduces the opportunity for ncRNA-mRNA crosstalk to impact translation. **(D)** Biophysical nucleation. Top left: a protoeukaryotic cell possessing endomembranes and protocoatomers (violet) that enable membrane bending (Devos et al., 2004). Bottom right: following a biophysical phase transition (Braun, 2008), the DNA becomes spontaneously encapsulated. The presence of protocoatomers prevents “catastrophic” encapsulation by forming protopores following membrane fusion. Protopores later diversify into nucleoporins (blue) and protocoatomers into coatomers (red, orange) (Devos et al., 2004).

limited under asexual reproduction (Poole, 2006). This model is relevant to eukaryote nuclear origins, but may not apply to other cases: the separation of transcription and translation (Gottshall et al., 2014) and presence of a nuclear envelope-like structure in *Gemmata* (Sagulenko et al., 2017) cannot be explained under this model, not least because *Gemmata* does not possess introns.

Avoidance of crosstalk may nevertheless be relevant to compartmentation. A recent study (Umu et al., 2016) demonstrated that there is selection for reduced crosstalk between highly expressed mRNAs and ncRNAs in a wide array of bacteria and archaea. Failure to avoid interaction with ncRNAs leads to reduced levels of protein expression, and hence it appears that such interactions have been selected against. For larger genomes with larger numbers of genes, it may be more challenging to avoid crosstalk interactions between ncRNAs and mRNAs. Moreover, where population sizes are also reduced (as expected in the origin of eukaryotes (Lynch, 2007)), this may reduce the capacity for crosstalk interactions to be selected against. Spatial separation of ncRNAs and mRNAs via compartmentation may thus alleviate crosstalk in eukaryotes, where weaker selection and larger numbers of transcriptional outputs would yield more opportunities for crosstalk. Indeed, in eukaryotes, where there is extensive RNA-based regulation, it is noteworthy that mRNA maturation (nucleus) and miRNA maturation (cytoplasm) are physically separated. Also suggestive of crosstalk avoidance, eukaryotic microRNAs are synthesized in a stem-loop precursor form (pre-miRNA) that, through internal base pairing, precludes crosstalk. Indeed, there is evidence for crosstalk avoidance in miRNAs during early embryonic development in *Drosophila* (Marco, 2015). Given that, at the very early stages of *Drosophila* development, multiple nuclear divisions occur in a syncytium without formation of cell membranes (Lawrence, 1992), the opportunities for crosstalk may be greater. For RNAi in general, it is only following export from the nucleus that the miRNA is processed by Dicer, and the interaction between mature miRNA and target mRNA occurs within the RNA-induced silencing complex (RISC) (Wilson and Doudna, 2013), which may again reduce the opportunities for cross-talk. Avoiding crosstalk may thus have been a factor in the evolution of the eukaryote nucleus, and possibly in bacteria with large genomes such as *G. obscuriglobus*.

In the case of jumbophage, avoidance might also be a driver of compartmentation, though for different reasons: it is tempting to consider that compartmentation might contribute to infection success by competitive exclusion of resident prophage.

Non-adaptive

A final possibility is that compartmentation, rather than being the result of natural selection, can in some cases non-adaptively emerge as a thermodynamic consequence of the amount of DNA present in a cell (Braun, 2008). In this model, colloidal phase separation in a crowded environment may occur if the genome is sufficiently large that the DNA nucleoid acts as a physical nucleation site (**Figure 2D**). Based on calculations that depend on intracellular volume and the number of expressed macromolecules, Braun (2008) estimated that formation of nucleoids may spontaneously occur for genomes that are ∼10 Mbp in length. In short, large genomes render the nascent growth of a compartment thermodynamically favorable. As appealing as this model is, it raises many questions. There are a few species of bacteria that have genomes ∼10 Mbp, and, notably, this does include *G. obscuriglobus* (∼9.2 Mbp). However, there are many bacteria that have very high genome copy number. For instance, the bacterium *Azotobacter vinlandii* has a chromosomal copy number exceeding 100 (Bendich and Drlica, 2000). This would push the total base-pair count to over an order of magnitude greater than Braun's calculations. Under this model, polyploid cells would need to be extremely large to avoid nucleation! Details aside, this model does not directly explain the formation of a membrane around the nucleoid, and lipid encapsulation may be fatal if transport is not already developed. However, it bears considering here as it raises the possibility that, rather than there being some direct advantage to compartmentalization, this might occur simply by biophysical processes. We should be wary of taking it as given that there was a strong selective advantage to nuclear compartmentation; it may instead have been non-lethal. Perhaps early coatomers (Devos et al., 2004) simply prevented complete encapsulation of DNA by invaginated membrane (**Figure 2D**), and, prior to the advent of FG-repeat proteins, larger molecules were afforded free movement between protonucleus and cytoplasm through the resulting proto-pores.

FUTURE PERSPECTIVES

The origins DNA compartmentation remains a difficult problem. However, the observation that this is not restricted to the eukaryote nucleus enables us to move away from the temptation of requiring some special explanation for eukaryote nuclear origins, to more mechanistic explanations, be they through neutral or selective processes. That there appear to be manifold routes to nuclear compartmentation is exciting not because different instances may be evolutionarily related, but rather because multiple instances may help us shed light on whether compartmentation of DNA is advantageous or accidental. In the context of the latter possibility, it is important to bear in mind that an accidental compartment might well turn out to be advantageous at some future date, without being of immediate short-term value; the key question for such a model is whether it is non-lethal in the short term. With ever-improving tools for synthetic biology and experimental evolution, it may soon be possible to start addressing these questions in the lab.

AUTHOR CONTRIBUTIONS

Both authors conceived and wrote the paper together.

FUNDING

AP thanks the Royal Society Te Apārangi, New Zealand, for support via a Rutherford Discovery Fellowship (RDF-11-UOC-013).

REFERENCES

Acehan, D., Santarella-Mellwig, R., and Devos, D. P. (2014). A bacterial tubulovesicular network. *J. Cell Sci.* 127, 277–280. doi: 10.1242/jcs.137596

Arkhipova, I., and Meselson, M. (2000). Transposable elements in sexual and ancient asexual taxa. *Proc. Natl. Acad. Sci. U.S.A.* 97, 14473–14477. doi: 10.1073/pnas.97.26.14473

Bapteste, E., Charlebois, R. L., MacLeod, D., and Brochier, C. (2005). The two tempos of nuclear pore complex evolution: highly adapting proteins in an ancient frozen structure. *Genome Biol.* 6:R85. doi: 10.1186/gb-2005-6-10-r85

Bell, P. J. (2001). Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? *J. Mol. Evol.* 53, 251–256. doi: 10.1007/s002390010215

Bell, P. J. (2009). The viral eukaryogenesis hypothesis: a key role for viruses in the emergence of eukaryotes from a prokaryotic world environment. *Ann. N. Y. Acad. Sci.* 1178, 91–105. doi: 10.1111/j.1749-6632.2009.04994.x

Bendich, A. J., and Drlica, K. (2000). Prokaryotic and eukaryotic chromosomes: what's the difference? *Bioessays* 22, 481–486.

Boedeker, C., Schüler, M., Reintjes, G., Jeske, O., van Teeseling, M. C., Jogler, M., et al. (2017). Determining the bacterial cell biology of Planctomycetes. *Nat. Commun.* 8:14853. doi: 10.1038/ncomms14853

Booth, A., and Doolittle, W. F. (2015). Eukaryogenesis, how special really? *Proc. Natl. Acad. Sci. U.S.A.* 112, 10278–10285. doi: 10.1073/pnas.1421376112

Braun, F. N. (2008). Thermodynamics of the prokaryote nuclear zone. *Int. J. Astrobiol.* 7, 183–185. doi: 10.1017/S1473550408004217

Chaikeeratisak, V., Nguyen, K., Egan, M. E., Erb, M. L., Vavilina, A., Pogliano, J., et al. (2017a). The phage nucleus and tubulin spindle are conserved among large *Pseudomonas* phages. *Cell Rep.* 20, 1563–1571. doi: 10.1016/j.celrep.2017.07.064

Chaikeeratisak, V., Nguyen, K., Khanna, K., Brilot, A. F., Erb, M. L., Coker, J. K., et al. (2017b). Assembly of a nucleus-like structure during viral replication in bacteria. *Science* 355, 194–197. doi: 10.1126/science.aal2130

Daly, M. J., Gaidamakova, E. K., Matrosova, V. Y., Vasilenko, A., Zhai, M., Leapman, R. D., et al. (2007). Protein oxidation implicated as the primary determinant of bacterial radioresistance. *PLoS Biol.* 5:e92. doi: 10.1371/journal.pbio.0050092

Davies, P. C., and Lineweaver, C. H. (2005). Finding a second sample of life on earth. *Astrobiology* 5, 154–163. doi: 10.1089/ast.2005.5.154

De Duve, C. (2005). *Singularities : Landmarks on the Pathways of Life*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511614736

Devos, D., Dokudovskaya, S., Alber, F., Williams, R., Chait, B. T., Sali, A., et al. (2004). Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* 2:e380. doi: 10.1371/journal.pbio.0020380

Dey, G., Thattai, M., and Baum, B. (2016). On the archaeal origins of eukaryotes and the challenges of inferring phenotype from genotype. *Trends Cell Biol.* 26, 476–485. doi: 10.1016/j.tcb.2016.03.009

Eme, L., Spang, A., Lombard, J., Stairs, C. W., and Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* 15, 711–723. doi: 10.1038/nrmicro.2017.133

Forterre, P., and Gaia, M. (2016). Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.* 31, 44–49. doi: 10.1016/j.mib.2016.02.001

Forterre, P., and Prangishvili, D. (2009). The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Ann. N. Y. Acad. Sci.* 1178, 65–77. doi: 10.1111/j.1749-6632.2009.04993.x

Fournier, G. P., and Poole, A. M. (2018). A briefly argued case that asgard archaea are part of the eukaryote tree. *Front. Microbiol.* 9:1896. doi: 10.3389/fmicb.2018.01896

Fuerst, J. A. (2005). Intracellular compartmentation in planctomycetes. *Annu. Rev. Microbiol.* 59, 299–328. doi: 10.1146/annurev.micro.59.030804.121258

Fuerst, J. A., and Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nat. Rev. Microbiol.* 9, 403–413. doi: 10.1038/nrmicro2578

Fuerst, J. A., and Sagulenko, E. (2012). Keys to eukaryality: planctomycetes and ancestral evolution of cellular complexity. *Front. Microbiol.* 3:167. doi: 10.3389/fmicb.2012.00167

Gabaldon, T. (2010). Peroxisome diversity and evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365, 765–773. doi: 10.1098/rstb.2009.0240

Gottshall, E. Y., Seebart, C., Gatlin, J. C., and Ward, N. L. (2014). Spatially segregated transcription and translation in cells of the endomembrane-containing bacterium *Gemmata obscuriglobus*. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11067–11072. doi: 10.1073/pnas.1409187111

Hendrix, R. W. (2009). Jumbo bacteriophages. *Curr. Top. Microbiol. Immunol.* 328, 229–240. doi: 10.1007/978-3-540-68618-7_7

Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* 101, 519–531.

Jeong, H., Sung, S., Kwon, T., Seo, M., Caetano-Anollés, K., Choi, S. H., et al. (2016). HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res.* 44, D610–D619. doi: 10.1093/nar/gkv1245

Klinger, C. M., Spang, A., Dacks, J. B., and Ettema, T. J. (2016). Tracing the archaeal origins of eukaryotic membrane-trafficking system building blocks. *Mol. Biol. Evol.* 33, 1528–1541. doi: 10.1093/molbev/msw034

Koonin, E. V., Makarova, K. S., and Wolf, Y. I. (2017). Evolutionary genomics of defense systems in Archaea and bacteria. *Annu. Rev. Microbiol.* 71, 233–261. doi: 10.1146/annurev-micro-090816-093830

Koumandou, V. L., Wickstead, B., Ginger, M. L., van der Giezen, M., Dacks, J. B., and Field, M. C. (2013). Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* 48, 373–396. doi: 10.3109/10409238.2013.821444

Krisko, A., and Radman, M. (2013). Biology of extreme radiation resistance: the way of *Deinococcus radiodurans*. *Cold Spring Harb. Perspect. Biol.* 5:a012765. doi: 10.1101/cshperspect.a012765

Kuznetsov, Y. G., Klose, T., Rossmann, M., and McPherson, A. (2013). Morphogenesis of mimivirus and its viral factories: an atomic force microscopy study of infected cells. *J. Virol.* 87, 11200–11213. doi: 10.1128/JVI.01372-13

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

Lane, N. (2011). Energetics and genetics across the prokaryote-eukaryote divide. *Biol. Direct* 6:35. doi: 10.1186/1745-6150-6-35

Lawrence, P. A. (1992). *The Making of a Fly*. Hoboken, NJ: Blackwell Scientific.

Lee, K. C., Webb, R. I., Janssen, P. H., Sangwan, P., Romeo, T., Staley, J. T., et al. (2009). Phylum Verrucomicrobia representatives share a compartmentalized cell plan with members of bacterial phylum Planctomycetes. *BMC Microbiol.* 9:5. doi: 10.1186/1471-2180-9-5

Leger, M. M., Eme, L., Stairs, C. W., and Roger, A. J. (2018). Demystifying eukaryote lateral gene transfer (Response to Martin 2017 doi: 10.1002/bies.201700115). *Bioessays* 40:e1700242. doi: 10.1002/bies.201700242

Lopez-Garcia, P., and Moreira, D. (2006). Selective forces for the origin of the eukaryotic nucleus. *Bioessays* 28, 525–533. doi: 10.1002/bies.20413

Lynch, M. (2007). *The Origins of Genome Architecture*. Sunderland, MA: Sinauer Associates.

Lynch, M., Bobay, L. M., Catania, F., Gout, J. F., and Rho, M. (2011). The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet.* 12, 347–366. doi: 10.1146/annurev-genom-082410-101412

Mans, B. J., Anantharaman, V., Aravind, L., and Koonin, E. V. (2004). Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* 3, 1612–1637. doi: 10.4161/cc.3.12.1316

Marco, A. (2015). Selection against maternal microRNA target sites in maternal transcripts. *G3* 5, 2199–2207. doi: 10.1534/g3.115.019497

Martin, W., Hoffmeister, M., Rotte, C., and Henze, K. (2001). An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. *Biol. Chem.* 382, 1521–1539. doi: 10.1515/BC.2001.187

Martin, W., and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440, 41–45. doi: 10.1038/nature04531

McInerney, J. O., Martin, W. F., Koonin, E. V., Allen, J. F., Galperin, M. Y., Lane, N., et al. (2011). Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays* 33, 810–817. doi: 10.1002/bies.201100045

McKay, C. P. (2004). What is life–and how do we search for it in other worlds? *PLoS Biol.* 2:E302. doi: 10.1371/journal.pbio.0020302

Mendoza, S. D., Berry, J. D., Nieweglowska, E. S., Leon, L. M., Agard, D., and Bondy-Denomy, J. (2018). A nucleus-like compartment shields bacteriophage DNA from CRISPR-Cas and restriction nucleases. *bioRxiv* [Preprint]. doi: 10.1101/370791

Mutsafi, Y., Shimoni, E., Shimon, A., and Minsky, A. (2013). Membrane assembly during the infection cycle of the giant *Mimivirus*. *PLoS Pathog.* 9:e1003367. doi: 10.1371/journal.ppat.1003367

Neumann, N., Lundin, D., and Poole, A. M. (2010). Comparative genomic evidence for a complete nuclear pore complex in the last eukaryote common ancestor. *PLoS One* 5:e13241. doi: 10.1371/journal.pone.0013241

Novoa, R. R., Calderita, G., Arranz, R., Fontana, J., Granzow, H., and Risco, C. (2005). Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biol. Cell* 97, 147–172. doi: 10.1042/BC20040058

Penny, D., Hendy, M. D., and Poole, A. M. (2003). Testing fundamental evolutionary hypotheses. *J. Theor. Biol.* 223, 377–385. doi: 10.1016/S0022-5193(03)00099-7

Poole, A., and Penny, D. (2007). Eukaryote evolution: engulfed by speculation. *Nature* 447:913. doi: 10.1038/447913a

Poole, A. M. (2006). Did group II intron proliferation in an endosymbiont-bearing archaeon create eukaryotes? *Biol. Direct* 1:36.

Race, H. L., Herrmann, R. G., and Martin, W. (1999). Why have organelles retained genomes? *Trends Genet.* 15, 364–370.

Rowley, P. A., Patterson, K., Sandmeyer, S. B., and Sawyer, S. L. (2018). Control of yeast retrotransposons mediated through nucleoporin evolution. *PLoS Genet.* 14:e1007325. doi: 10.1371/journal.pgen.1007325

Sagulenko, E., Morgan, G. P., Webb, R. I., Yee, B., Lee, K.-C., and Fuerst, J. A. (2014). Structural studies of planctomycete *Gemmata obscuriglobus* support cell compartmentalisation in a bacterium. *PLoS One* 9:e91344. doi: 10.1371/journal.pone.0091344

Sagulenko, E., Nouwens, A., Webb, R. I., Green, K., Yee, B., Morgan, G., et al. (2017). Nuclear pore-like structures in a compartmentalized bacterium. *PLoS One* 12:e0169432. doi: 10.1371/journal.pone.0169432

Sinninghe Damste, J. S., Strous, M., Rijpstra, W. I., Hopmans, E. C., Geenevasen, J. A., van Duin, A. C., et al. (2002). Linearly concatenated cyclobutane lipids form a dense bacterial membrane. *Nature* 419, 708–712. doi: 10.1038/nature01128

Slade, D., Lindner, A. B., Paul, G., and Radman, M. (2009). Recombination and replication in DNA repair of heavily irradiated *Deinococcus radiodurans*. *Cell* 136, 1044–1055. doi: 10.1016/j.cell.2009.01.018

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., et al. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521, 173–179. doi: 10.1038/nature14447

Staley, J. T., and Fuerst, J. A. (2017). Ancient, highly conserved proteins from a LUCA with complex cell biology provide evidence in support of the nuclear compartment commonality (NuCom) hypothesis. *Res. Microbiol.* 168, 395–412. doi: 10.1016/j.resmic.2017.01.001

Suzan-Monti, M., La Scola, B., Barrassi, L., Espinosa, L., and Raoult, D. (2007). Ultrastructural characterization of the giant volcano-like virus factory of *Acanthamoeba polyphaga Mimivirus*. *PLoS One* 2:e328. doi: 10.1371/journal.pone.0000328

Takemura, M. (2001). Poxviruses and the origin of the eukaryotic nucleus. *J. Mol. Evol.* 52, 419–425. doi: 10.1007/s002390010171

Timney, B. L., Raveh, B., Mironska, R., Trivedi, J. M., Kim, S. J., Russel, D., et al. (2016). Simple rules for passive diffusion through the nuclear pore complex. *J. Cell Biol.* 215, 57–76. doi: 10.1083/jcb.201601004

Umu, S. U., Poole, A. M., Dobson, R. C., and Gardner, P. P. (2016). Avoidance of stochastic RNA interactions can be harnessed to control protein expression levels in bacteria and archaea. *eLife* 5:e13479. doi: 10.7554/eLife.13479

van Niftrik, L. (2013). Cell biology of unique anammox bacteria that contain an energy conserving prokaryotic organelle. *Antonie Van Leeuwenhoek* 104, 489–497. doi: 10.1007/s10482-013-9990-5

Werren, J. H. (2011). Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 2), 10863–10870. doi: 10.1073/pnas.1102343108

Wilson, R. C., and Doudna, J. A. (2013). Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* 42, 217–239. doi: 10.1146/annurev-biophys-083012-130404

Yuan, Y., and Gao, M. (2017). Jumbo bacteriophages: an overview. *Front. Microbiol.* 8:403. doi: 10.3389/fmicb.2017.00403

Zaremba-Niedzwiedzka, K., Caceres, E. F., Saw, J. H., Bäckström, D., Juzokaite, L., Vancaester, E., et al. (2017). Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541, 353–358. doi: 10.1038/nature21031

Conflict of Interest Statement: 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.

REVIEW
published: 29 January 2019
doi: 10.3389/fmicb.2019.00051

Evolution of Immune Systems From Viruses and Transposable Elements

Felix Broecker[1]* and *Karin Moelling*[2,3]

[1] Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, United States, [2] Institute of Medical Microbiology, University of Zurich, Zurich, Switzerland, [3] Max Planck Institute for Molecular Genetics, Berlin, Germany

Virus-derived sequences and transposable elements constitute a substantial portion of many cellular genomes. Recent insights reveal the intimate evolutionary relationship between these sequences and various cellular immune pathways. At the most basic level, superinfection exclusion may be considered a prototypical virus-mediated immune system that has been described in both prokaryotes and eukaryotes. More complex immune mechanisms fully or partially derived from mobile genetic elements include CRISPR-Cas of prokaryotes and the RAG1/2 system of vertebrates, which provide immunological memory of foreign genetic elements and generate antibody and T cell receptor diversity, respectively. In this review, we summarize the current knowledge on the contribution of mobile genetic elements to the evolution of cellular immune pathways. A picture is emerging in which the various cellular immune systems originate from and are spread by viruses and transposable elements. Immune systems likely evolved from simple superinfection exclusion to highly complex defense strategies.

Keywords: transposable elements, mobile genetic elements, viruses, superinfection exclusion, immune system, CRISPR-Cas, antibodies, RNase H

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana-Champaign, United States

Reviewed by:
Dino McMahon,
Freie Universität Berlin, Germany
Alan Richard Davidson,
University of Toronto, Canada

****Correspondence:***
Felix Broecker
felixbroecker@gmx.net

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

Received: *04 November 2018*
Accepted: *14 January 2019*
Published: *29 January 2019*

Citation:
Broecker F and Moelling K (2019)
Evolution of Immune Systems From
Viruses and Transposable Elements.
Front. Microbiol. 10:51.
doi: 10.3389/fmicb.2019.00051

INTRODUCTION

Cellular organisms have co-evolved with various mobile genetic elements (MGEs), including transposable elements (TEs), retroelements and viruses, many of which can integrate into the host DNA. MGEs constitute ∼50% of mammalian genomes, >70% of some plant genomes and up to 30% of bacterial genomes (Koonin and Krupovic, 2015). The evolutionary interplay between MGEs and their hosts has generated a plethora of cellular defense mechanisms and counter-measures. Notably, many immune systems, or parts thereof, including the prokaryotic CRISPR-Cas mechanism and antibody/T cell receptor (TCR) diversification by V(D)J recombination in vertebrates have been recruited from viruses or other MGEs. Here, we summarize the current knowledge on the evolution of diverse immune systems of prokaryotes and eukaryotes, highlighting a general scenario for the origin of cellular defense systems from MGEs. A non-exhaustive overview of different cellular immune systems is presented in **Figure 1**.

VIRUSES AGAINST VIRUSES: SUPERINFECTION EXCLUSION AS A MECHANISM OF ANTIVIRAL IMMUNITY

Superinfection exclusion (SIEx) is the ability of a preexisting viral infection to restrict secondary infections, often by the same or a closely related virus. SIEx was first observed in tobacco plants that, when pre-infected with a mild variant of *Tobacco mosaic virus* (TMV), were protected against a virulent TMV strain (McKinney, 1929). SIEx was later found to be common for many other systems, including viruses of bacteria, animals, humans, and plants (Moelling et al., 2017). The cellular organism benefits from SIEx if a preexisting infection with a non-pathogenic or mildly pathogenic virus protects against detrimental viruses. Thus, SIEx can be regarded as a simple adaptive immune system, which is inheritable if the first virus integrates into the cellular genome or is transmitted to the progeny by other means. One recent experimentally verified example is Mavirus, a virophage that integrates into the genome of *Cafeteria roenbergensis* and protects the flagellate organism from infection with a deadly virus (Fischer and Hackl, 2016). This example is further described below.

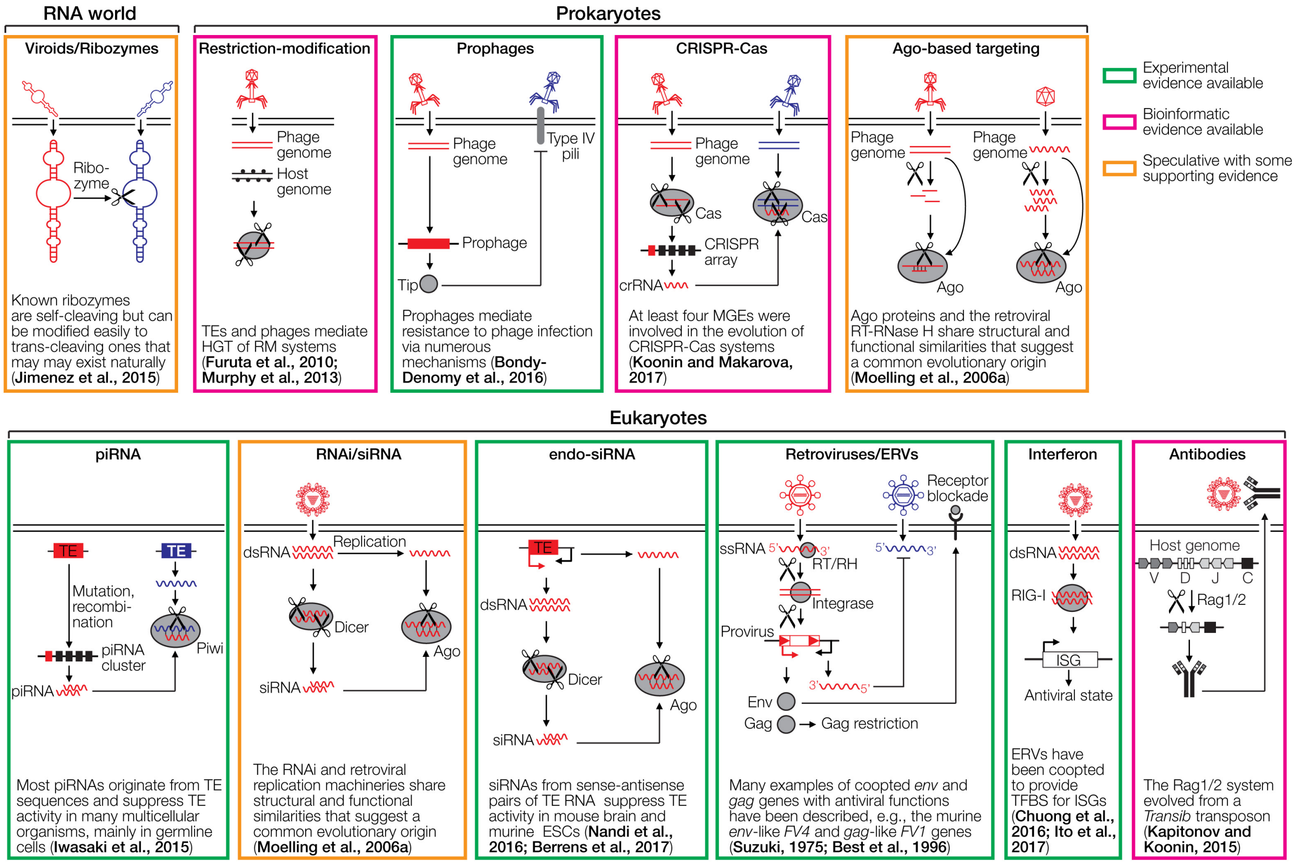
An evolutionarily early immune system may have been constituted by viroids or viroid-like RNAs. Viroids are virus-related, protein-free infectious agents consisting of highly structured, circular non-coding RNA that can be catalytically active through ribozyme activity (Flores et al., 2014). They may be remnants of the ancient RNA world thought to have existed before the evolution of DNA or proteins (Diener, 1989; Flores et al., 2014). However, the fact that extant viroids have so far only been identified in plants (with the notable exception of hepatitis delta virus, a derivative of a viroid with a short insert of protein-coding capacity) suggests their appearance after the last universal cellular ancestor (Koonin and Dolja, 2014). Regardless, viroids likely recapitulate principal features of selfish elements of the ancient RNA world. In plants, SIEx has been described between mild and severe strains of the same viroid as well as between different viroids (Kovalskaya and Hammond, 2014). The mechanisms of SIEx in plants may include RNA interference (RNAi), with siRNAs produced by Dicer from the first infecting viroid acting against the superinfecting one. It remains unclear, however, how the first viroid escapes RNAi; it may associate with protecting host factors or its localization in the nucleus or chloroplasts protects from RNAi, which mainly acts in the cytoplasm (Kovalskaya and Hammond, 2014). It seems likely that SIEx existed before the evolution of complex viruses or cellular immune systems such as RNAi. In the ancient RNA world, a simple RNA-based immune system could have been constituted of a ribozyme/viroid that prevents superinfection with another one *via* ribozymatic cleavage *in trans* (**Figure 1**). Although known natural ribozymes/viroids are generally self-cleaving, they can be modified relatively easily to yield *trans*-cleaving derivatives (Jimenez et al., 2015), suggesting that *trans*-cleaving ribozymes may have existed or may still exist naturally.

INTEGRATED VIRAL SEQUENCES ACT AS INHERITABLE IMMUNITY IN PROKARYOTES

Insertion of viral genomes, or parts thereof, into host genomes is at the origin of many immune systems (Moelling et al., 2017). Integration of prophages into bacterial genomes is often associated with a fitness cost to the host (Iranzo et al., 2017), however, prophages can mediate resistance to infection by exogenous bacteriophages (phages) through various mechanisms. For example, the prophage-encoded Tip protein inhibits formation of type IV pili on the surface of *Pseudomonas aeruginosa* (Chung et al., 2014). Since these pili are common phage receptors, Tip expression mediates SIEx to various phages (Bondy-Denomy et al., 2016). Interestingly, prophage-mediated alteration of type IV pili function has little or no fitness cost to the host. In *P. aeruginosa*, three prophages are sufficient to mediate resistance against at least 30 different phages. Various other mechanisms of prophage-mediated protection against exogenous phages have been reported in multiple bacterial species, which include cell surface alterations, receptor blockade and transcriptional repression (Bondy-Denomy and Davidson, 2014).

CRISPR-Cas provides another prokaryotic adaptive immune system. Here, a fragment of DNA (or reverse-transcribed RNA) of an infecting phage or other foreign genetic elements is integrated as spacer into a CRISPR array in the host genome (Hille et al., 2018). Thereby, spacers act as immunological memory for long-term protection of the cell and future generations. The transcribed CRISPR array RNA (pre-crRNA) is processed into smaller crRNAs that guide sequence-specific cleavage of homologous invading nucleic acids by Cas effector nucleases. A common feature of all Type II and Type V effector nucleases, Cas9 and Cas12a, respectively, is a ribonuclease H (RNase H)-like RuvC domain that cleaves the non-complementary DNA strand, whereas an HNH nuclease (Cas9) or a NUC domain (Cas12a) cleaves the complementary DNA strand of the target dsDNA (Makarova et al., 2017). Other Cas effectors utilize different nuclease domains, such as histidine-aspartate nucleases. CRISPR arrays are found in about half of bacterial and nearly 90% of archaeal genomes (Hille et al., 2018).

At least four different MGEs were involved in the evolution of CRISPR-Cas systems (Koonin and Makarova, 2017). First, the adaptation module of all CRISPR-Cas systems, responsible for spacer acquisition, originated from casposons (a fusion word between Cas and transposons), TEs that utilize Cas1 nuclease for DNA integration (Krupovic et al., 2014). Second, the Cas2 nuclease, which could have been present also in the casposon, as well as HEPN family RNases found in several other Cas proteins likely originate from toxin-antitoxin (TA) modules (Koonin and Krupovic, 2015). Although TA modules do not encode their own mobility genes, they can be regarded as MGEs, as they are typically transferred by plasmids (Koonin and Makarova, 2017). Third, many Type III CRISPR-Cas systems recruited the reverse transcriptase (RT) from a mobile group II intron, which allows for spacer acquisition from invading RNA. Fourth, the RuvC domains of Type II and Type V systems were likely derived from TnpB nucleases of DNA transposons. A complete, functional CRISPR-Cas system encoded by a phage has been reported (Seed et al., 2013), suggesting that phages may also serve as vehicles for horizontal gene transfer (HGT) of this kind of immune system. Both prophage-mediated SIEx and CRISPR-Cas can be regarded as adaptive prokaryotic immune systems that generate immunological memory to protect the cell and future generations against viral infections. CRISPR-Cas acquired specific immunity can be transmitted



Known ribozymes are self-cleaving but can be modified easily to trans-cleaving ones that may may exist naturally (**Jimenez et al., 2015**)

TEs and phages mediate HGT of RM systems (**Furuta et al., 2010; Murphy et al., 2013**)

Prophages mediate resistance to phage infection via numerous mechanisms (**Bondy-Denomy et al., 2016**)

At least four MGEs were involved in the evolution of CRISPR-Cas systems (**Koonin and Makarova, 2017**)

Ago proteins and the retroviral RT-RNase H share structural and functional similarities that suggest a common evolutionary origin (**Moelling et al., 2006a**)

Most piRNAs originate from TE sequences and suppress TE activity in many multicellular organisms, mainly in germline cells (**Iwasaki et al., 2015**)

The RNAi and retroviral replication machineries share structural and functional similarities that suggest a common evolutionary origin (**Moelling et al., 2006a**)

siRNAs from sense-antisense pairs of TE RNA suppress TE activity in mouse brain and murine ESCs (**Nandi et al., 2016; Berrens et al., 2017**)

Many examples of coopted *env* and *gag* genes with antiviral functions have been described, e.g., the murine *env*-like *FV4* and *gag*-like *FV1* genes (**Suzuki, 1975; Best et al., 1996**)

ERVs have been coopted to provide TFBS for ISGs (**Chuong et al., 2016; Ito et al., 2017**)

The Rag1/2 system evolved from a *Transib* transposon (**Kapitonov and Koonin, 2015**)

FIGURE 1 | Cartoons depicting various defense systems. The systems are color-coded based on the level of support in green (experimental evidence available), magenta (bioinformatic evidence available), and orange (speculative with some supporting evidence). DNA/RNA cleaving is indicated with scissors. RNA is depicted as wavy lines or with secondary hairpin-loop structures. A ribozyme cleaving another ribozyme is a hypothetical early immune system that does not require DNA or proteins. The ribozyme may be part of a viroid-like selfish RNA. Restriction-modification systems distinguish between foreign and self DNA by methylating target sequences of restriction endonucleases. Prophages can mediate superinfection exclusion, exemplified by expression of the Tip protein that reduces bacterial surface expression of type IV pili required for infection by various phages. CRISPR-Cas acts by incorporating small genomic fragments from phages into CRISPR arrays in the prokaryotic genome. The transcribed spacers are then used by another Cas member to cleave sequence-homologous phage genomes. PIWI-associated RNAs (piRNAs) are small RNAs complementary to transposable elements (TEs) that are encoded in piRNA clusters (Iwasaki et al., 2015). piRNAs associate with a PIWI nuclease to cleave complementary TE transcripts. RNA interference (RNAi) is initiated by dsRNA which is fragmented by Dicer to siRNAs. These siRNAs are loaded into the RNA-induced silencing complex to cleave complementary RNAs using Ago nucleases. A variation of RNAi is the endo-siRNA pathway, in which dsRNA is generated from TEs that are transcribed in both orientations, for instance, if the TE is located in an intron in opposite orientation to the encompassing gene. Endogenous retroviruses (ERVs) can mediate restriction of ERVs and exogenous retroviruses through various mechanisms, including receptor blockade by captured Env proteins, Gag-mediated restriction and antisense RNA mechanisms. The interferon system recognizes dsRNA or other pathogen-associated molecular patterns, which leads to upregulation of antiviral interferon-stimulated genes (ISGs). The antibody system involves diversification through light and heavy chain recombination, which is mediated by the Rag recombinases. This enables the detection of diverse pathogens. Note that the references provided in the figure are not comprehensive. Please refer to the text for more details and additional references.

across thousands of microbial generations (Weinberger et al., 2012).

BACTERIAL IMMUNE SYSTEMS CO-LOCALIZE WITH VIRAL AND TRANSPOSABLE ELEMENT SEQUENCES

Restriction-modification (RM) systems consist of two components; a restriction endonuclease that cleaves invading double-stranded DNA (e.g., of phage origin) by recognizing a short DNA motif and a methylase that masks that motif on the prokaryote's genome by introducing methyl groups to prevent destruction of its own DNA. Since the recognition motifs are usually short and thereby likely to be present on the majority of invading DNA molecules, RM systems can be regarded as a prokaryotic innate immune system. The various restriction endonucleases of RM systems likely evolved from one or a few common ancestor(s) (Jeltsch et al., 1995) and became widespread *via* HGT (Jeltsch and Pingoud, 1996). RM systems are encoded by about 90% of prokaryotes (Murphy et al., 2013). Various phages have been shown to be able to mediate HGT of RM genes, indicating that phages are common vectors for these immune systems (Murphy et al., 2013). RM genes frequently co-localize with viral and TE sequences such as integrases and transposases and in some cases are flanked by inverted repeats and target site duplications, hallmarks of TEs (Naderer et al., 2002; Furuta et al., 2010; Makarova et al., 2011; Takahashi et al., 2011). TEs carrying functional RM systems have been identified (Khan et al., 2010), raising the possibility that these defense systems evolutionarily originate from TEs. Some restriction endonucleases can also trigger programmed cell death of bacteria (Nagamalleswari et al., 2017). This phenomenon of 'bacterial apoptosis' has been described as a mechanism that occurs upon phage infection to limit spread of the virus, reminiscent of eukaryotic apoptosis triggered by viral infection (Chopin et al., 2005).

A number of additional prokaryotic innate anti-phage systems have recently been identified (Koonin, 2018). These include prokaryotic Ago proteins that cleave invading DNA or RNA with RNase H-like nuclease domains (Swarts et al., 2014), BREX, which blocks phage replication and methylates bacterial DNA, enabling BREX to differentiate between host and phage genomes (Goldfarb et al., 2015) and DISARM, which also methylates host DNA and restricts invading dsDNA phages (Ofir et al., 2017). In addition, a number of defense systems were recently identified by a systematic search for genes clustering with defense islands, regions involved in defense processes that are widely abundant in prokaryotic genomes (Makarova et al., 2011; Doron et al., 2018). Ten of the novel defense systems were verified experimentally *in vitro* either in *Escherichia coli* or *Bacillus subtilis* that became resistant to a panel of phages upon introduction of the defense system. Interestingly, TE sequences are enriched in defense islands (Doron et al., 2018). The majority of prokaryotic TEs encode DDE superfamily transposases with an RNase H fold (named after two aspartate and one glutamate residue that form a catalytic triad) that mediate mobility *via* a cut-and-paste mechanism (Koonin and Krupovic, 2015). It remains unknown if these TEs serve a functional role, or whether their accumulation in defense island is simply less deleterious compared to other genomic loci. It is tempting to speculate, however, that some of the defense island-associated TE sequences, especially the RNase H-like transposases, may have been captured by the host to fulfill defense functions, or that they have contributed to the spread of immune systems.

PROTECTION FROM RETROVIRAL INFECTION BY ENDOGENIZED *env* GENES

Eukaryotic genomes harbor large amounts of endogenous retrovirus (ERV) sequences, which are remnants of retroviral infections of ancestral germline cells. The human and mouse genomes, for instance, contain about 8% and 10% ERV sequences, respectively (Gifford and Tristem, 2003; Broecker et al., 2016). Among the best studied examples of retroviral genes that have been captured by mammalian (and some reptilian) hosts are the syncytins (Lavialle et al., 2013). Syncytins originate from *env* genes of endogenized proviruses. Full-length proviruses harbor the three retroviral genes, *gag*, *pol* and *env*, flanked by two LTRs. ERVs not subject to any selective pressure are inactivated by mutation to various degrees over time (Broecker et al., 2016). In rare occasions, however, certain proviral genes have been conserved over millions of years of evolution (**Figure 2A**), suggesting a selective advantage of that gene to the host. Retroviral *env* genes have been repeatedly captured from different proviruses at least 17 times during evolution and as syncytins or related genes exert critical physiological functions in the placental development of various mammalian and viviparous lizard species (**Figure 2B**) (Cornelis et al., 2017; Imakawa and Nakagawa, 2017).

Syncytins also likely contribute to maternal immune tolerance toward the fetus *via* the immunosuppressive domain (ISD) (**Figure 2C**). The ISD has been demonstrated to exert various immunosuppressive functions *in vitro* and *in vivo*, including an inhibition of the activity of lymphocytes, natural killer cells, monocytes and macrophages as well as the downregulation of pro-inflammatory cytokines (Cianciolo et al., 1985; Haraguchi et al., 1995, 1997, 2008). An *env*-derived syncytin gene has recently been identified in viviparous lizards that possess a mammalian-like placenta (Cornelis et al., 2017). Thus, syncytin capture is not restricted to mammals and likely a hallmark of placental evolution in general.

In addition to syncytins and related genes, other retroviral *env* genes have been captured for anti-retroviral functions. The first described example, the *Friend virus susceptibility 4* (*Fv4*) gene, confers resistance to murine leukemia viruses (MuLVs) in mice (Suzuki, 1975). *FV4* is a truncated MuLV-related provirus containing the 3′ portion of *pol*, the entire *env* gene and the 3′LTR (Ikeda et al., 1985). Binding of the *env*-encoded protein to the cellular receptor used by MuLV prevents infection by the exogenous retrovirus, a process termed receptor blockade. Another captured *env* gene in mice, *resistance to MCF* (*Rmcf*), mediates resistance to mink cell focus-inducing (MCF) viruses and MuLVs, likely also *via* receptor blockade (Hartley et al., 1983; Brightman et al., 1991; Jung et al., 2002).

Jaagsiekte sheep retrovirus (JSRV) co-exists both as exogenous and endogenous form (Armezzani et al., 2014). JSRV is an example of recent or ongoing endogenization, with the youngest identified endogenous elements (enJSRV) having integrated about 200 years ago. The sheep genome harbors at least 27 enJSRV sequences, 16 of which with intact *env* genes. enJSRV *env* is expressed in the ovine placenta and knockdown with antisense oligonucleotides has been shown to cause defects in trophoblast differentiation and pregnancy loss (Dunlap et al., 2006). This demonstrates that even recently endogenized, mostly intact *env* genes can exert syncytin-like functions. In addition, enJSRV *env* expression has been shown to block exogenous JSRV *via* receptor blockade (Spencer et al., 2003).

In cats, the *refrex-1* gene mediates resistance to feline leukemia virus-D (FeLV-D) (Ito et al., 2013). *Refrex-1* is a truncated

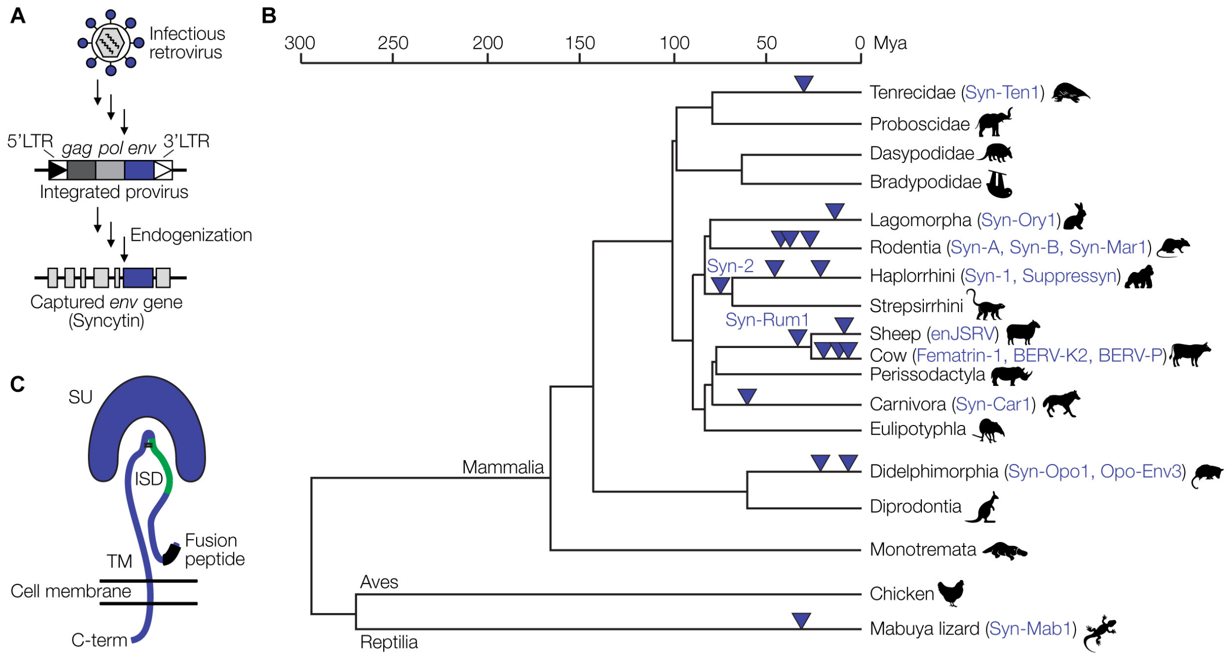


FIGURE 2 | Capture of retroviral genes for placental development. **(A)** Schematic of endogenization and capture of a syncytin gene. A retrovirus infecting a germline cell integrates into the host genome as a provirus with characteristic features, the 5′LTR, *gag*, *pol* and *env* genes, and the 3′LTR. Over time, in a process termed endogenization, most of the provirus acquires deleterious mutations. In this example, the *env* gene retains an intact open reading frame and is captured by the host to fulfill functions during placental development. **(B)** Evolution of placental species and capture of ERV-derived *env* genes. The phylogenetic tree is based on previously published data (Cornelis et al., 2017; Imakawa and Nakagawa, 2017). **(C)** Structural representation of the retroviral Env protein with the surface (SU), transmembrane (TM), and immunosuppressive domain (ISD) subunits, as well as the fusion peptide. Panel **(C)** has been modified from Lavialle et al. (2013).

retroviral *env* gene, such that the protein contains the signal peptide (SP) and the N-terminus of the surface unit (SU) that is a putative receptor-binding domain, but lacks the C-terminus of SU and the transmembrane (TM) domain due to a premature stop codon (**Figure 2C**).

In human cells, it has been recently shown that the Env protein encoded by a HERV-K(HML-2) provirus interferes with HIV-1 production *in vitro* through an unknown mechanism (Terry et al., 2017). The provirus, HERV-K108, encodes full-length Env with four mutations compared to the consensus, ancestral HERV-K(HML-2) protein. These mutations appear to be required for inhibiting HIV-1. Interestingly, HERV-K(HML-2) expression in T cells is increased upon HIV-1 infection (Contreras-Galindo et al., 2007; Gonzalez-Hernandez et al., 2012). It is therefore tempting to speculate that inducible HERV-K(HML-2) proviruses have been evolutionarily conserved to express Env (and Gag, see below) to protect against exogenous retroviruses such as HIV. Another example of an evolutionarily conserved *env* gene with antiviral function in human cells is HERV-T *env* (Blanco-Melo et al., 2017). Expression of this gene mediates resistance to a reconstructed infectious HERV-T virus (the virus is extinct) by receptor blockade. Interestingly, the consensus *env* of the resurrected virus, but not the single endogenized *env* gene could be used by the virus for successful infection. This indicates that the *env* gene has been modified evolutionarily to bind to the viral receptor while losing its ability to constitute infectious virions. In addition, *Suppressyn*, a truncated *env* gene from a HERV-F element with a known role in placental development (see above), has been suggested to serve as restriction factor for exogenous retroviruses (Malfavon-Borja and Feschotte, 2015).

PROTECTION FROM RETROVIRAL INFECTION BY ENDOGENIZED *gag* GENES

Another retroviral gene that has been frequently captured by mammalian hosts is *gag*. The best studied *gag*-derived restriction factor is the mouse gene *Friend virus susceptibility 1* (*Fv1*) (Best et al., 1996). *Fv1* inhibits MuLV at a stage post-entry but before integration of the provirus, with the exact mechanism still unknown. The Fv1 protein has been shown to interact with the retroviral capsid protein (CA) in the preintegration complex of MuLV (Best et al., 1996). *Fv1* originates from a MERV-L *gag* gene that has little sequence homology with that of MuLV, implying that Fv1 and its target share structural properties despite few sequence similarities. In sheep, enJSRV expressed Gag has been shown to inhibit virion formation of exogenous JSRV, however, in contrast to Fv1, at a late stage during viral assembly (Palmarini et al., 2004). In human cells, the HERV-K(HML-2) CA protein inhibits release of HIV-1 and reduces infectivity of progeny HIV-1 virions (Monde et al., 2017).

OTHER ANTIVIRAL EFFECTS MEDIATED BY ENDOGENOUS RETROVIRUSES

In addition to Env- and Gag-mediated restriction, more indirect mechanisms of antiviral protection by ERVs have been described. In human cells, the HERV-K Rec protein expressed during early embryogenesis activates innate immune responses by inducing expression of the *IFITM1* gene, which protects the cell from viral infection (Grow et al., 2015). Moreover, ERVs have introduced and amplified interferon (IFN)-inducible enhancers within eukaryotic genomes and provide transcription factor binding sites (TFBS) that are enriched in proximity to genes involved in immune pathways (Chuong et al., 2016; Ito et al., 2017). This suggests that ERV sequences have been specifically adopted by host cells to modulate IFN responses, a major branch of the antiviral immune defense. An example in the human genome is the HERV-K(HML-10) family recently described by us and others (Broecker et al., 2016; Grandi et al., 2017). TFBS within the LTRs are frequently occupied, as determined by the ENCODE project (Davis et al., 2018), especially in the human myelogenous leukemia cell line K562 (**Figure 3A**). HERV-K(HML-10) elements appear to be enriched in loci involved in immunity such as the extended major histocompatibility complex (xMHC) and the extended leukocyte receptor complex (xLRC)

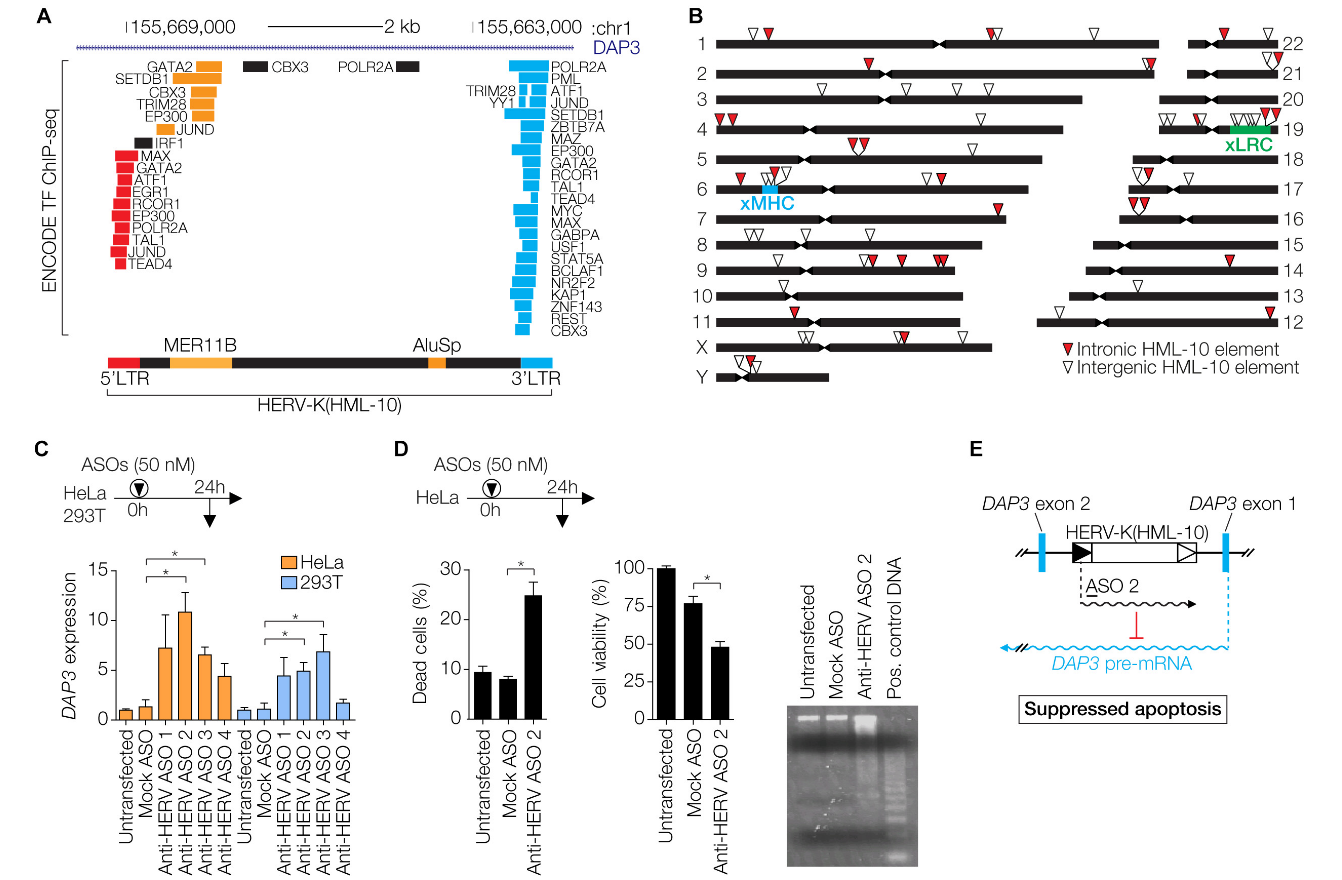


FIGURE 3 | Regulatory functions of HERV-K(HML-10) elements. **(A)** Genomic region of a HERV-K(HML-10) provirus in the first intron of the *DAP3* gene. Repeat elements are according to RepeatMasker (Smit et al., 2013–2015) annotation. Two retroelements, *MER11B* and *AluSp*, are integrated into the provirus. Occupied transcription factor binding sites, as determined by ENCODE (Davis et al., 2018), are indicated and color-coded according to localization within the HERV-K(HML-10) provirus. Image modified from the UCSC Genome Browser (Kent et al., 2002) with the hg19 release of the human genome. **(B)** Chromosomal distribution of HERV-K(HML-10) elements in the human genome. The locations of the extended major histocompatibility complex (xMHC) and the extended leukocyte receptor complex (xLRC) are indicated. **(C)** Cells were transfected with indicated antisense oligonucleotides (ASOs). Anti-HERV ASOs target a HERV-K(HML-10) derived regulatory transcript, Mock ASO has an unrelated sequence. At 24 h after transfection, *DAP3* mRNA expression levels were determined by qRT-PCR and normalized to *GAPDH* levels. Bars show mean ± SEM of three experiments, untransfected cells were set to 1. *$P \leq 0.05$, Student's *t*-test against Mock. **(D)** HeLa cells were transfected with the indicated ASOs. After 48 h, Trypan Blue exclusion as indicator of dead cells (left) and MTS cell viability assays (center) were performed. The right subpanel shows genomic DNA of these cells prepared with the Apoptotic DNA Ladder Kit (Roche) and analyzed on an ethidium bromide stained agarose gel. The positive control DNA is from apoptotic U937 cells provided with the kit. Bars show mean ± SEM of three experiments in duplicates. *$P \leq 0.05$, Student's *t*-test. **(E)** Model of regulation of *DAP3* expression by the HERV-K(HML-10) primed regulatory transcript. The approximate location of ASO 2 is indicated. This figure shows data modified from Broecker et al. (2016).

(Barrow and Trowsdale, 2008; Horton et al., 2004) (**Figure 3B**). The data suggests that HML-10 has been captured by the host for the regulation of immune-related genes.

We have recently described a HERV-K(HML-10) provirus that regulates *DAP3* gene expression *in vitro* through an antisense mechanism, likely *via* a long non-coding RNA (**Figure 3C**) (Broecker et al., 2016). Inactivating the HERV-derived RNA by antisense oligonucleotides was sufficient to induce apoptosis *in vitro* (**Figure 3D**), demonstrating that HERV-mediated antisense regulation can directly influence the cellular phenotype (**Figure 3E**). In addition, there is evidence for HERV-mediated gene regulation in humans *in vivo*. The complement component *C4* gene in the xMHC exists in two variants, one with a transcriptionally active intron-located HERV-K(HML-10) provirus, and one without it (Schneider et al., 2001; Yang et al., 2003; Mack et al., 2004; Broecker et al., 2016). The presence of the provirus correlates with lower C4 protein serum concentrations, indicating that the HERV regulates *C4* expression, perhaps *via* an antisense mechanism (Yang et al., 2003). The fact that HERV-derived RNA can regulate expression of cellular genes also suggests that HERV antisense transcripts arising from intronic proviruses could suppress mRNA of exogenous retroviruses. In support of this, HERV antisense transcripts arising from intron-located proviruses have been shown to suppress complementary retroviral transcripts *in trans* (Schneider et al., 2001; Mack et al., 2004). Recombination between exogenous retroviruses and ERVs is another restriction mechanism, which may occur when defective ERV transcripts are co-packaged with the intact retroviral RNA into the same virion, or at the level of proviral DNA during meiosis (Löber et al., 2018).

ENDOGENIZATION IN REAL TIME: KOALA RETROVIRUS

An ongoing retroviral endogenization occurs in koalas. Koala retrovirus (KoRV) co-exists as both exogenous and endogenous form (Stoye, 2006; Tarlinton et al., 2006). High KoRV viral loads are associated with fatal lymphoid neoplasia. Since integration sites and copy numbers of proviruses are heterogeneous among individuals, and some koala populations isolated from mainland Australia since around the year 1900 are free of KoRV, the virus has likely entered the koala genome only about 100 years ago and is still undergoing endogenization. KoRV endogenization is possibly associated with resistance of koalas against the exogenous virus (Colson et al., 2015). To date it is unclear if Env or Gag-mediated restriction mechanisms may protect from exogenous KoRV. However, many endogenous proviruses have *gag* and *env* genes with complete open reading frames (ORFs), usually with point mutations (Oliveira et al., 2007). The mutated KoRV genes, when incorporated into retroviral vectors based on the closely related gibbon ape leukemia virus (GALV) that uses the same entry receptor as KoRV substantially reduced infectivity compared to the GALV *gag* and *env* genes. This suggests that KoRV endogenization is associated with mutations that render the proviruses incapable of producing highly infectious viruses. Yet, preservation of intact *gag* and *env* ORFs may indicate a functional importance, perhaps as restriction factors against exogenous KoRV. Moreover, some koala populations such as those located in Southern Australia (SA) are relatively resistant to KoRV induced disease and usually have low viral loads (Tarlinton et al., 2017). In contrast, koalas in northern Australia such as Queensland suffer from higher viremia and disease burden, and active KoRV infection is more prevalent. Of note, both populations have endogenous KoRV proviruses which, however, differ in their RNA expression patterns. While koalas from Queensland express mainly full-length proviruses, including the complete *gag*, *pol* and *env* genes, koalas from SA preferentially express the 5′ portion of *gag* and the 3′ portion of *env*, whereas *pol* transcripts are weak or undetectable. This may reflect ongoing endogenization events in which short variants of Gag and Env are preserved that mediate restriction to exogenous KoRV. Of note, the Fv1 restriction factor in mice is also not a full-length Gag protein but covers the first three-fourth of the retroviral Gag protein it is likely derived from (Bénit et al., 1997). Another explanation for why koalas from Queensland are not protected from KoRV is that expression of the full-length KoRV genes tolerizes the animals to the virus *in utero*, rendering their immune systems unable to recognize and respond to the exogenous virus. This may explain why many animals from Queensland do not have antibodies to KoRV and also do not elicit them upon vaccination (Fiebig et al., 2015; Waugh et al., 2016).

HIV *EN ROUTE* TO ENDOGENIZATION?

Is HIV currently ongoing endogenization in humans? HIV is a complex retrovirus of the *lentivirus* genus. In contrast, ERVs are typically derived from simple retroviruses. The first endogenous lentivirus, RELIK (rabbit endogenous lentivirus-K) was identified only in 2007 (Katzourakis et al., 2007) [simple ERVs have been known since the 1960s (Weiss, 2006)], followed by the reports of several endogenous lentiviruses in primate (Gifford et al., 2008; Gilbert et al., 2009; Han and Worobey, 2015), ferret (Cui and Holmes, 2012), and weasel genomes (Han and Worobey, 2012). Still, the vast majority of known ERVs are derived from simple retroviruses, suggesting that endogenization of lentiviruses has been a relatively rare event, perhaps due to the recent evolutionary origin of lentiviruses (Katzourakis et al., 2007).

A prerequisite for endogenization to occur is the capability of a retrovirus to infect germline cells. Whether HIV can infect human germline cells is controversial. HIV-1 particles have been detected on the cell membrane and inside *in vitro* infected spermatozoa (Baccetti et al., 1994, 1998; Barboza et al., 2004; Cardona-Maya et al., 2009). In addition, HIV-1 proviral DNA has been detected by PCR in spermatozoa of HIV-1 infected individuals (Bagasra et al., 1994; Nuovo et al., 1994; Cardona-Maya et al., 2009). HIV-1 particles can associate with spermatozoa by binding to mannose receptor, which can transfer the virus into oocytes (Baccetti et al., 1994; Cardona-Maya et al., 2011). These findings indicate that sperm cells can be infected and subject to provirus integration, and can also act as vectors of HIV-1 virions to infect oocytes. Therefore, endogenization of HIV-1 appears to be possible in theory. Vertical transmission of proviral DNA through the germline, however, has not been demonstrated yet.

A recently reported HIV-1 patient controlling the infection without antiretroviral therapy despite not having the CCR5-Δ32 mutation or a protective HLA genotype suggests that the presence of HIV-1 proviruses in lymphocytes may protect against AIDS (Colson et al., 2014). PBMCs from this patient harbored defective HIV-1 proviruses and could not be superinfected with the same strain of HIV-1 *in vitro*, suggesting that the proviruses rendered the PBMCs resistant to infection. The proviruses harbored a number of premature stop codons likely introduced by the APOBEC3G restriction factor, however, some ORF were intact. The presence of apparently protective HIV-1 proviruses suggests that SIEx mediated by HIV-1 proviruses is likely possible. Thus, a potential germline infection with HIV-1 may confer inheritable resistance against HIV-1 induced disease.

PROTECTION AGAINST VIRAL INFECTION BY ENDOGENIZED NON-RETROVIRAL GENES IN EUKARYOTES

Mammalian genomes not only contain ERVs and TEs, but also a number of sequences derived from *Bornaviridae*, *Filoviridae*, *Parvoviridae*, *Circoviridae*, *Rhabdoviridae*, and others (Belyi et al., 2010a,b; Horie et al., 2010; Katzourakis and Gifford, 2010; Aswad and Katzourakis, 2014). Genomic sequences from RNA viruses without an RT are likely processed pseudogenes originating from illegitimate reverse transcription and integration by the replication machinery of long interspersed nuclear elements (LINEs), or they have arisen from recombination with ERV RNA (Suzuki et al., 2014).

The best studied example of the function of non-retroviral endogenous virus sequences are Borna disease viruses (BDVs). BDVs are neurotropic negative sense ssRNA viruses causing fatal encephalitis (Borna disease) in horses, sheep and cattle (Carbone, 2001; Belyi et al., 2010b). These highly susceptible species have no detectable endogenous BDV sequences in their genomes (Belyi et al., 2010b). BDV also persistently infects other species, from avian to primate, and in experimental animals such as mice can induce behavioral changes without obvious signs of encephalitis. Interestingly, the genomes of primates, rats, and mice and squirrels, which are relatively resistant to the virus, harbor BDV sequences (Belyi et al., 2010b). This suggests a protective role of endogenous BDV sequences in protecting against disease caused by exogenous BDV.

Experimental evidence for protection mediated by endogenous BDV sequences has been obtained in squirrels. The genome of the 13-lined ground squirrel *Ictidomys tridecemlineatus* contains an endogenous bornavirus-like nucleoprotein (itEBLN) sequence that shares 77% amino acid similarity with current infectious BDV (Fujino et al., 2014). itEBLN colocalizes with the viral factory of BDV in the nucleus and suppresses viral replication and cell-to-cell spread *in vitro*, likely acting as a dominant negative nucleoprotein that is incorporated into BDV virions, which renders them non-infectious. Thus, itEBLN may serve as immune memory against exogenous BDV.

Like squirrels, humans usually do not experience Borna disease, with only three cases of fatal BDV-induced viral encephalitis reported to date resulting from zoonotic infections from squirrels (Hoffmann et al., 2015) and three more cases (two of which fatal) of human-to-human transmission during organ transplantation (Friedrich-Loeffler-Institut, 2018). All of the seven human endogenous bornavirus-like nucleoprotein elements (hsEBLN-1 through hsEBLN-7) are expressed as RNA in one or more tissues (Sofuku et al., 2015). At least one of them, hsEBLN-2, is also expressed as protein in human cells (Ewing et al., 2007). In primates and rodents, EBLNs are significantly enriched in piRNA clusters (Parrish et al., 2015). Three of the seven hsEBLN genes and three of five rodent EBLNs are located in piRNA clusters. Interestingly, piRNA cluster-located EBLNs in both rodents and primates produce bona fide piRNAs, which are antisense relative to the BDV nucleoprotein mRNA and are expressed in the testes. Whether piRNA-mediated inhibition of BDV infection occurs in germline cells, however, remains to be determined. As piRNAs can also be expressed in somatic cells including neurons (Lee et al., 2011), EBLN sequences may also protect from BDV infection in the brain, which may at least partially explain the resistance of species with endogenous EBLN sequences to viral encephalitis. Another potential mechanism by which EBLNs may protect from Borna disease is the induction of immune tolerance by *in utero* expression of EBLN protein (Horie, 2017). Tolerization of the adaptive immune system to the EBLN protein *in utero* may limit the immune response to the nucleoprotein during BDV infection. The BDV nucleoprotein is known as a major target for cytotoxic T cell responses (Planz and Stitz, 1999). *In utero* tolerization to this antigen may be protective, as most of the symptoms of fatal BDV infection arise from immune-mediated inflammation. A further antiviral mechanism might occur at the RNA level. EBLN RNAs could act as antisense transcripts to the genomic minus sense ssRNA genome of BDV that replicates in the nucleus (Horie, 2017). Aside from squirrels, primates and rodents, EBLN sequences have been identified in Afrotherians, bats, whales, birds, and lamprey (Kobayashi et al., 2016; Hyndman et al., 2018). It is conceivable that some of these elements also exert anti-viral functions.

The genomes of *Aedes* mosquitoes which are important vectors for human pathogenic flaviviruses such as Dengue and Zika contain various endogenous flaviviral sequences (Suzuki et al., 2017). piRNAs and siRNAs are produced from these endogenous viruses and might play a role in antiviral defense. It is known that small RNAs play an important role in antiviral defense in insects (Cullen et al., 2013).

A variation of SIEx can also be mediated by viruses that infect other viruses, termed virophages. The protozoan *Cafeteria roenbergensis* is infected by *Cafeteria roenbergensis* virus (CroV), a giant virus that causes lysis of the host (Fischer and Suttle, 2011). CroV is infected by the virophage Mavirus (Fischer and Hackl, 2016). *C. roenbergensis* cells co-infected with CroV and Mavirus are protected from lysis, as Mavirus inhibits replication of CroV. Interestingly, Mavirus can integrate into the genome of *C. roenbergensis* where it stays inactive until the cell is infected with CroV. The activated Mavirus then inhibits CroV replication, thus providing an adaptive, inducible immunity of *C. roenbergensis* against detrimental CroV infection. Stably integrated into the *C. roenbergensis* genome, Mavirus is passed on to the next generation of the protozoan, which can be regarded as a simple form of an inheritable immune system.

Polintons, TEs related to virophages, are found in the genomes of diverse eukaryotic species and likely originate from viruses with an exogenous form (Koonin and Krupovic, 2018). They may represent endogenized virophages that, unlike Mavirus in *C. roenbergensis*, have lost the ability to form virions. Polintons may have been recruited by their eukaryotic hosts as a defense against past or present viruses, whose identity remains to be determined.

ADAPTIVE IMMUNITY OF JAWED VERTEBRATES: V(D)J RECOMBINATION

In contrast to the adaptive immune system of prokaryotes, CRISPR-Cas, immunological memory in vertebrates is restricted to somatic cells and is therefore not inherited to the next generation. In jawed vertebrates, the diversity of immunoglobulins/antibodies and TCRs is generated by V(D)J recombination, in which variable (V), diversity (D) and joining (J) segments are recombined. Further antibody diversification is then achieved by somatic hypermutation (Kapitonov and Koonin, 2015).

The ability to produce diversity of antibodies and TCRs in jawed vertebrates developed at around 500 mya (Kapitonov and Koonin, 2015). Both the Rag1 and Rag2 proteins required for V(D)J recombination are found in one genomic locus and originate from a *Transib* transposon that today is found in the genomes of starfish, oysters and sea urchins, but not anymore in those of jawed vertebrates (Kapitonov and Koonin, 2015). Rag1 is the nuclease responsible for V(D)J recombination, which contains an RNase H-like domain with the conserved DDE catalytic triad.

POSSIBLE EVOLUTION OF COMPONENTS OF RNA INTERFERENCE FROM VIRUSES

The retroviral replication machinery and Argonaute (Ago)-mediated interference pathways against invading nucleic acids share some surprising similarities at the structural and functional level (Moelling et al., 2006a, 2017; Moelling and Broecker, 2015). At the core of retroviral replication is the reverse transcriptase-RNase H (RT-RNase H) that resembles Ago proteins consisting of PAZ (N-terminal), MID (central) and PIWI (C-terminal) domains (**Figure 4A**). The C-terminal domains of both proteins adopt an RNase H-fold, one of the most ancient and abundant protein folds found in nature (**Figures 4B,C**) (Wang et al., 2006; Ma et al., 2008; Majorek et al., 2014). These domains have nuclease activity and cleave the target RNA/DNA (the template retroviral RNA during retroviral replication or the nucleic acid targeted by Ago *via* the guide nucleic acid) (Mölling et al., 1971; Song et al., 2004). The N-terminal domains of both proteins, RT of the RT-RNase H and PAZ of Ago, among other functions, serve as nucleic acid binding modules that direct the cleavage specificity of the RNase H domains. The RNA/DNA binding activity in the RT domain of RT-RNase H is located in conserved residues binding the template RNA strand ("template grip") as well as the opposite cDNA strand ("primer grip") (Dash et al., 2008). In the case of Ago, the PAZ domain is an oligonucleotide-binding domain that interacts with the 3′ end of the guide (Song et al., 2004). Further interactions with the 5′ end of the guide are made by the MID domain (Lima et al., 2009). In both RT-RNase H and Ago, the N-terminal domains fused to the RNase H domain determine the specificity of the nuclease activity. The widespread presence of Ago proteins in prokaryotes and eukaryotes with conserved structures and functions suggests an ancient evolutionary origin, possibly before the last eukaryotic common ancestor (Swarts et al., 2014). The diverse Ago proteins can act as RNA- or DNA-guided nucleases and can cleave RNA or DNA through the RNase H-like PIWI domain. In both prokaryotes and eukaryotes, Ago-centered defense can be regarded as a mechanism of innate immunity (Koonin and Krupovic, 2015).

RNAi is triggered by siRNAs, short (21–24 nucleotides) double-stranded RNAs with two nucleotide overhangs at the 3′ ends, which are produced by Dicer from longer dsRNA molecules. A DNA molecule structurally related to siRNAs, a partially double-stranded hairpin-loop 54-mer oligodeoxynucleotide (ODN) is an efficient inducer of the HIV-1 RT-RNase H, leading to cleavage of the retroviral RNA genome (Moelling et al., 2006b). Interestingly, the human AGO2 protein can use both siRNAs and ODNs to find and cleave target RNA in a sequence-specific manner *in vitro* – albeit with lower efficiency with the ODN (Moelling et al., 2006a). *Vice versa*, siRNAs are recognized by the HIV-1 RT-RNase H to induce target RNA cleavage but with lower efficiency than the corresponding ODN. These common activities suggest an evolutionary relationship between the RT-RNase H and AGO2. Another component required for RNAi, the RNA-dependent RNA polymerase (RdRP) that amplifies siRNAs, likely originated from a phage (Shabalina and Koonin, 2008). The RdRP was likely present in the last eukaryotic common ancestor and is still active in plants and nematodes (Shabalina and Koonin, 2008).

While eukaryotic Ago proteins are generally believed to only use RNA as guides, prokaryotic Ago proteins have been demonstrated previously to also accept DNA as guides, with important functions *in vivo* (Yuan et al., 2005). The abovementioned activity of AGO2 with a DNA guide against HIV-1 RNA in the test tube suggests that mammalian RNAi may also be triggered by DNA guides to serve biological functions. Indeed, DNA molecules have been shown to bind to the PAZ domain of AGO2 and localize into mRNA-degrading P bodies, hallmark features of RNAi-mediated degradation (Castanotto

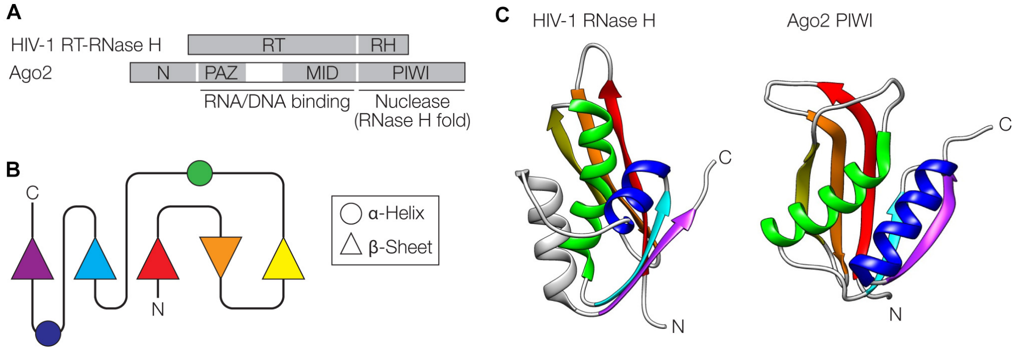


FIGURE 4 | Similarities between retroviral reverse-transcriptase (RT)-RNase H and Argonaute. **(A)** Domain organization of the HIV-1 RT-RNase H and Argonaute 2 (Ago2) from *Pyrococcus furiosus*. The RT domain of RT-RNase H and the Piwi Argonaute Zwille (PAZ) and middle (MID) domains of Ago2 contain regions involved in nucleic acid binding. The RNase H (RH) and PIWI domains adopt an RNase H fold and serve as nucleases. **(B)** Schematic of the RNase H fold. **(C)** The RNase H and PIWI domains are shown next to each other, with α-helices and β-sheets color-coded as in panel **(B)**. PDB accession numbers are 2I5J for the RNase H (Himmel et al., 2006) and 1U04 for Ago2 (Song et al., 2004) The gray helix in the HIV structure is an additional helix not shown in **(B)**. Structures were visualized with UCSF Chimera (Pettersen et al., 2004).

et al., 2015). It has been suggested that cytosolic genomic DNA (cgDNA) functions as natural antisense mechanism triggering RNA degradation, perhaps involving AGO2 (Asada et al., 2018). Single-stranded cgDNA of TE origin can be detected in mammalian cell lines and may act as a natural antisense mechanism against the RNA of retrotransposons, especially ERVs (Stetson et al., 2008; Asada et al., 2018). However, under normal conditions *in vivo* the exonuclease TREX1 appears to degrade single-stranded cgDNA, preventing the antisense inhibition. Notably, loss of function mutations in the human *TREX1* gene cause Aicardi-Goutières Syndrome (Crow et al., 2006), an autoimmune disease characterized by the accumulation of cytosolic ssDNA (Yang et al., 2007), including ssDNA of TEs, especially of ERV origin (Stetson et al., 2008). Thus, cgDNA may restrict TE transcripts *via* RNAi under conditions without or with low TREX1 expression.

Another mechanism by which TE-derived nucleic acids lead to inhibition of TE expression is the endogenous siRNA (endosiRNA) pathway (Ghildiyal et al., 2008; Nandi et al., 2016; Berrens et al., 2017). In this pathway, TE sense-antisense RNA pairs that arise, for instance, from intron-located TEs (**Figure 1**) are subject to RNAi, involving Dicer and AGO2, which suppresses TE activity in the mammalian germline (Berrens et al., 2017), the mammalian brain (Nandi et al., 2016) as well in somatic cells of *Drosophila* (Ghildiyal et al., 2008). It appears possible that a similar siRNA-based mechanism may also act against exogenous retroviral RNA if there is sufficient complementarity between an antisense-transcribed ERV and the mRNA of the exogenous retrovirus (**Figure 1**). Indeed, mammalian RNAi has been demonstrated *in vitro* to also act against exogenous viruses. This includes RNAi-mediated restriction of enteroviruses (Qiu et al., 2017), encephalomyocarditis virus and Nodamura virus (Maillard et al., 2013), Influenza virus (Matskevich and Moelling, 2007; Li et al., 2016), reovirus and Sindbis virus (Maillard et al., 2016). The biological relevance of siRNA in mammalian cells, however, is subject to debate, as many mammalian viruses efficiently counteract RNAi (tenOever, 2017; Tsai et al., 2018). Moreover, mammalian antiviral RNAi is usually only detected in cells defective in IFN signaling and may be restricted to embryonic stem cells. In plants, nematodes and invertebrates, however, RNAi plays an important role in antiviral defense (Cullen et al., 2013).

CONCLUSION

Recruitment of sequences from viruses, TEs, and other MGEs for immune defense mechanisms in prokaryotes and eukaryotes is strikingly common. Nucleases are involved in many immune systems, either to cleave invading DNA or to mediate genome editing events (**Figure 1**). Many of these are RNase H-like nucleases, including Ago/Piwi proteins involved in foreign nucleic acid cleavage in prokaryotes and in RNAi in eukaryotes, some CRISPR-Cas effector nucleases (Cas9 and Cas12), and the Rag1 protein that mediates V(D)J recombination. Thus, RNase H-like molecules are involved in different prokaryotic and eukaryotic immune systems of various origin. The fact that genomes of almost all cellular organisms harbor large numbers of MGEs suggests that yet unknown functionalities may be identified in the future. The recent discovery that defense islands in bacterial genomes are enriched with sequences of TEs further pinpoints their important role in immune defense mechanisms in prokaryotes. Their potential functions remain to be elucidated. In eukaryotes whose genomes usually contain even more TE sequences than prokaryotic ones, additional immune functions are also expected to be discovered. This includes TE-derived conserved genes such as *HARBI1*, which is conserved across vertebrates and originated from a Harbinger transposase, with yet unknown functions (Koonin and Krupovic, 2015).

It has to be noted that antiviral defense is by far not the only function of endogenized viruses and TEs. For example, deletion of all replication-deficient prophages in *E. coli* has resulted in fitness deficits under diverse environmental conditions, including increased susceptibility to antibiotics and osmotic stress, slower cell growth and reduced biofilm formation (Wang et al., 2010). In eukaryotes, TEs and ERVs do not only modulate IFN response genes and constitute antiviral defense mechanisms, but also play distinct roles in cell differentiation, stem cell pluripotency and embryonic development, amongst others (Chuong et al., 2017), and the industrial melanism mutation of peppered moths has been shown to be caused by a TE insertion (Van't Hof et al., 2016). These examples highlight the multifaceted roles of TEs and viral sequences in pro- and eukaryotes. However, given their diverse roles in various immune systems (**Figure 1**), it appears that recruitment of TEs, viral sequences and other MGEs for antiviral defense mechanisms has been a major driving force in the evolution of cellular life.

AUTHOR CONTRIBUTIONS

Both authors have written the manuscript and approved its final version.

FUNDING

FB was financially supported by a fellowship of the German Academy of Sciences Leopoldina.

ACKNOWLEDGMENTS

We would like to thank Prof. Peter Palese for his generous support.

REFERENCES

Armezzani, A., Varela, M., Spencer, T. E., Palmarini, M., and Arnaud, F. (2014). "Ménage à Trois": the evolutionary interplay between JSRV, enJSRVs and domestic sheep. *Viruses* 6, 4926–4945. doi: 10.3390/v6124926

Asada, K., Ito, K., Yui, D., Tagaya, H., and Yokota, T. (2018). Cytosolic genomic DNA functions as a natural antisense. *Sci. Rep.* 8:8551. doi: 10.1038/s41598-018-26487-1

Aswad, A., and Katzourakis, A. (2014). The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. *PLoS Genet.* 10:e1004332. doi: 10.1371/journal.pgen.1004332

Baccetti, B., Benedetto, A., Burrini, A. G., Collodel, G., Ceccarini, E. C., Crisà, N., et al. (1994). HIV-particles in spermatozoa of patients with AIDS and their transfer into the oocyte. *J. Cell Biol.* 127, 903–914. doi: 10.1083/jcb.127.4.903

Baccetti, B., Benedetto, A., Collodel, G., di Caro, A., Garbuglia, A. R., and Piomboni, P. (1998). The debate on the presence of HIV-1 in human gametes. *J. Reprod. Immunol.* 41, 41–67. doi: 10.1016/S0165-0378(98)00048-5

Bagasra, O., Farzadegan, H., Seshamma, T., Oakes, J. W., Saah, A., and Pomerantz, R. J. (1994). Detection of HIV-1 proviral DNA in sperm from HIV-1-infected men. *AIDS* 8, 1669–1674. doi: 10.1097/00002030-199412000-00005

Barboza, J. M., Medina, H., Doria, M., Rivero, L., Hernandez, L., and Joshi, N. Y. (2004). Use of atomic force microscopy to reveal sperm ultrastructure in HIV-patients on highly active antiretroviral therapy. *Arch. Androl.* 50, 121–129.

Barrow, A. D., and Trowsdale, J. (2008). The extended human leukocyte receptor complex: diverse ways of modulating immune responses. *Immunol. Rev.* 224, 98–123. doi: 10.1111/j.1600-065X.2008.00653.x

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010a). Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old. *J. Virol.* 84, 12458–12462. doi: 10.1128/JVI.01789-10

Belyi, V. A., Levine, A. J., and Skalka, A. M. (2010b). Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. *PLoS Pathog.* 6:e1001030. doi: 10.1371/journal.ppat.1001030

Bénit, L., De Parseval, N., Casella, J. F., Callebaut, I., Cordonnier, A., and Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J. Virol.* 71, 5652–5657.

Berrens, R. V., Andrews, S., Spensberger, D., Santos, F., Dean, W., Gould, P., et al. (2017). An endosiRNA-based repression mechanism counteracts transposon activation during global DNA demethylation in embryonic stem cells. *Cell Stem Cell* 21, 694–703. doi: 10.1016/j.stem.2017.10.004

Best, S., Le Tissier, P., Towers, G., and Stoye, J. P. (1996). Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* 382, 826–829. doi: 10.1038/382826a0

Blanco-Melo, D., Gifford, R. J., and Bieniasz, P. D. (2017). Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *eLife* 6:e22519. doi: 10.7554/eLife.22519

Bondy-Denomy, J., and Davidson, A. R. (2014). When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *J. Microbiol.* 52, 235–242. doi: 10.1007/s12275-014-4083-3

Bondy-Denomy, J., Qian, J., Westra, E. R., Buckling, A., Guttman, D. S., Davidson, A. R., et al. (2016). Prophages mediate defense against phage infection through diverse mechanisms. *ISME J.* 10, 2854–2866. doi: 10.1038/ismej.2016.79

Brightman, B. K., Li, Q. X., Trepp, D. J., and Fan, H. (1991). Differential disease restriction of Moloney and Friend murine leukemia viruses by the mouse Rmcf gene is governed by the viral long terminal repeat. *J. Exp. Med.* 174, 389–396. doi: 10.1084/jem.174.2.389

Broecker, F., Horton, R., Heinrich, J., Franz, A., Schweiger, M. R., Lehrach, H., et al. (2016). The intron-enriched HERV-K(HML-10) family suppresses apoptosis, an indicator of malignant transformation. *Mob. DNA* 7:25. doi: 10.1186/s13100-016-0081-9

Carbone, K. M. (2001). Borna disease virus and human disease. *Clin. Microbiol. Rev.* 14, 513–527. doi: 10.1128/CMR.14.3.513-527.2001

Cardona-Maya, W., Velilla, P., Montoya, C. J., Cadavid, A., and Rugeles, M. T. (2009). Presence of HIV-1 DNA in spermatozoa from HIV-positive patients: changes in the semen parameters. *Curr. HIV Res.* 7, 418–424. doi: 10.2174/157016209788680570

Cardona-Maya, W., Velilla, P. A., Montoya, C. J., Cadavid, Á., and Rugeles, M. T. (2011). In vitro human immunodeficiency virus and sperm cell interaction mediated by the mannose receptor. *J. Reprod. Immunol.* 92, 1–7. doi: 10.1016/j.jri.2011.09.002

Castanotto, D., Lin, M., Kowolik, C., Wang, L., Ren, X. Q., Soifer, H. S., et al. (2015). A cytoplasmic pathway for gapmer antisense oligonucleotide-mediated gene silencing in mammalian cells. *Nucleic Acids Res.* 43, 9350–9361. doi: 10.1093/nar/gkv964

Chopin, M. C., Chopin, A., and Bidnenko, E. (2005). Phage abortive infection in lactococci: variations on a theme. *Curr. Opin. Microbiol.* 8, 473–479. doi: 10.1016/j.mib.2005.06.006

Chung, I. Y., Jang, H. J., Bae, H. W., and Cho, Y. H. (2014). A phage protein that inhibits the bacterial ATPase required for type IV pilus assembly. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11503–11508. doi: 10.1073/pnas.1403537111

Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351, 1083–1087. doi: 10.1126/science.aad5497

Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71–86. doi: 10.1038/nrg.2016.139

Cianciolo, G. J., Copeland, T. D., Oroszlan, S., and Snyderman, R. (1985). Inhibition of lymphocyte proliferation by a synthetic peptide homologous to retroviral envelope proteins. *Science* 230, 453–455. doi: 10.1126/science.2996136

Colson, P., Levy, Y., and Raoult, D. (2015). Response to a letter to the editor by Joachim Denner on HIV infection en route to endogenization: two cases. *Clin. Microbiol. Infect.* 21, e35–e37. doi: 10.1016/j.cmi.2014.11.029

Colson, P., Ravaux, I., Tamalet, C., Glazunova, O., Baptiste, E., Chabriere, E., et al. (2014). HIV infection en route to endogenization: two cases. *Clin. Microbiol. Infect.* 20, 1280–1288. doi: 10.1111/1469-0691.12807

Contreras-Galindo, R., López, P., Vélez, R., and Yamamura, Y. (2007). HIV-1 infection increases the expression of human endogenous retroviruses type K (HERV-K) in vitro. *AIDS Res. Hum. Retrovi.* 23, 116–122. doi: 10.1089/aid.2006.0117

Cornelis, G., Funk, M., Vernochet, C., Leal, F., Tarazona, O. A., Meurice, G., et al. (2017). An endogenous retroviral envelope syncytin and its cognate receptor identified in the viviparous placental *Mabuya* lizard. *Proc. Natl. Acad. Sci. U.S.A.* 114, E10991–E11000. doi: 10.1073/pnas.1714590114

Crow, Y. J., Leitch, A., Hayward, B. E., Garner, A., Parmar, R., Griffith, E., et al. (2006). Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. *Nat. Genet.* 38, 910–916. doi: 10.1038/ng1842

Cui, J., and Holmes, E. C. (2012). Endogenous lentiviruses in the ferret genome. *J. Virol.* 86, 3383–3385. doi: 10.1128/JVI.06652-11

Cullen, B. R., Cherry, S., and tenOever, B. R. (2013). Is RNA interference a physiologically relevant innate antiviral immune response in mammals? *Cell Host Microbe* 14, 374–378. doi: 10.1016/j.chom.2013.09.011

Dash, C., Scarth, B. J., Badorrek, C., Götte, M., and Le Grice, S. F. (2008). Examining the ribonuclease H primer grip of HIV-1 reverse transcriptase by charge neutralization of RNA/DNA hybrids. *Nucleic Acids Res.* 36, 6363–6371. doi: 10.1093/nar/gkn678

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Diener, T. O. (1989). Circular RNAs: relics of precellular evolution? *Proc. Natl. Acad. Sci. U.S.A.* 86, 9370–9374. doi: 10.1073/pnas.86.23.9370

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., et al. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* 359:eaar4120. doi: 10.1126/science.aar4120

Dunlap, K. A., Palmarini, M., Varela, M., Burghardt, R. C., Hayashi, K., Farmer, J. L., et al. (2006). Endogenous retroviruses regulate periimplantation placental growth and differentiation. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14390–14395. doi: 10.1073/pnas.0603836103

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3:89. doi: 10.1038/msb4100134

Fiebig, U., Keller, M., Möller, A., Timms, P., and Denner, J. (2015). Lack of antiviral antibody response in koalas infected with koala retroviruses (KoRV). *Virus Res.* 198, 30–34. doi: 10.1016/j.virusres.2015.01.002

Fischer, M. G., and Hackl, T. (2016). Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* 540, 288–291. doi: 10.1038/nature20593

Fischer, M. G., and Suttle, C. A. (2011). A virophage at the origin of large DNA transposons. *Science* 332, 231–234. doi: 10.1126/science.1199412

Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., and Elena, S. F. (2014). Viroids: survivors from the RNA world? *Annu. Rev. Microbiol.* 68, 395–414. doi: 10.1146/annurev-micro-091313-103416

Friedrich-Loeffler-Institut. (2018). Humane infektionen mit dem borna disease virus (BoDV-1) [in German]. *Epid. Bull.* 10:105.

Fujino, K., Horie, M., Honda, T., Merriman, D. K., and Tomonaga, K. (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13175–13180. doi: 10.1073/pnas.1407046111

Furuta, Y., Abe, K., and Kobayashi, I. (2010). Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res.* 38, 2428–2443. doi: 10.1093/nar/gkp1226

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., et al. (2008). Endogenous siRNAs derived from transposons and mRNAs in *Drosophila* somatic cells. *Science* 320, 1077–1081. doi: 10.1126/science.1157396

Gifford, R., and Tristem, M. (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 26, 291–315. doi: 10.1023/A:1024455415443

Gifford, R. J., Katzourakis, A., Tristem, M., Pybus, O. G., Winters, M., and Shafer, R. W. (2008). A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20362–20367. doi: 10.1073/pnas.0807873105

Gilbert, C., Maxfield, D. G., Goodman, S. M., and Feschotte, C. (2009). Parallel germline infiltration of a lentivirus in two Malagasy lemurs. *PLoS Genet.* 5:e1000425. doi: 10.1371/journal.pgen.1000425

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., et al. (2015). BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* 34, 169–183. doi: 10.15252/embj.201489455

Gonzalez-Hernandez, M. J., Swanson, M. D., Contreras-Galindo, R., Cookinham, S., King, S. R., Noel, R. J. Jr., et al. (2012). Expression of human endogenous retrovirus type K (HML-2) is activated by the Tat protein of HIV-1. *J. Virol.* 86, 7790–7805. doi: 10.1128/JVI.07215-11

Grandi, N., Cadeddu, M., Pisano, M. P., Esposito, F., Blomberg, J., and Tramontano, E. (2017). Identification of a novel HERV-K(HML10): comprehensive characterization and comparative analysis in non-human primates provide insights about HML10 proviruses structure and diffusion. *Mob. DNA* 8:15. doi: 10.1186/s13100-017-0099-7

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522, 221–225. doi: 10.1038/nature14308

Han, G. Z., and Worobey, M. (2012). Endogenous lentiviral elements in the weasel family (Mustelidae). *Mol. Biol. Evol.* 29, 2905–2908. doi: 10.1093/molbev/mss126

Han, G. Z., and Worobey, M. (2015). A primitive endogenous lentivirus in a colugo: insights into the early evolution of lentiviruses. *Mol. Biol. Evol.* 32, 211–215. doi: 10.1093/molbev/msu297

Haraguchi, S., Good, R. A., Cianciolo, G. J., Engelman, R. W., and Day, N. K. (1997). Immunosuppressive retroviral peptides: immunopathological implications for immunosuppressive influences of retroviral infections. *J. Leukoc. Biol.* 61, 654–666. doi: 10.1002/jlb.61.6.654

Haraguchi, S., Good, R. A., and Day-Good, N. K. (2008). A potent immunosuppressive retroviral peptide: cytokine patterns and signaling pathways. *Immunol. Res.* 41, 46–55. doi: 10.1007/s12026-007-0039-6

Haraguchi, S., Good, R. A., James-Yarish, M., Cianciolo, G. J., and Day, N. K. (1995). Differential modulation of Th1- and Th2-related cytokine mRNA expression by a synthetic peptide homologous to a conserved domain within retroviral envelope protein. *Proc. Natl. Acad. Sci. U.S.A.* 92, 3611–3615. doi: 10.1073/pnas.92.8.3611

Hartley, J. W., Yetter, R. A., and Morse, H. C. III (1983). A mouse gene on chromosome 5 that restricts infectivity of mink cell focus-forming recombinant murine leukemia viruses. *J. Exp. Med.* 158, 16–24. doi: 10.1084/jem.158.1.16

Hille, F., Richter, H., Wong, S. P., Bratoviè, M., Ressel, S., and Charpentier, E. (2018). The biology of CRISPR-Cas: backward and forward. *Cell* 172, 1239–1259. doi: 10.1016/j.cell.2017.11.032

Himmel, D. M., Sarafianos, S. G., Dharmasena, S., Hossain, M. M., McCoy-Simandle, K., Ilina, T., et al. (2006). HIV-1 reverse transcriptase structure with RNase H inhibitor dihydroxy benzoyl naphthyl hydrazone bound at a novel site. *ACS Chem. Biol.* 1, 702–712. doi: 10.1021/cb600303y

Hoffmann, B., Tappe, D., Höper, D., Herden, C., Boldt, A., Mawrin, C., et al. (2015). A variegated squirrel bornavirus associated with fatal human encephalitis. *N. Engl. J. Med.* 373, 154–162. doi: 10.1056/NEJMoa1415627

Horie, M. (2017). The biological significance of bornavirus-derived genes in mammals. *Curr. Opin. Virol.* 25, 1–6. doi: 10.1016/j.coviro.2017.06.004

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., et al. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463, 84–87. doi: 10.1038/nature08695

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., et al. (2004). Gene map of the extended human MHC. *Nat. Rev. Genet.* 5, 889–899. doi: 10.1038/nrg1489

Hyndman, T. H., Shilton, C. M., Stenglein, M. D., and Wellehan, J. F. X. Jr. (2018). Divergent bornaviruses from Australian carpet pythons with neurological disease date the origin of extant *Bornaviridae* prior to the end-Cretaceous extinction. *PLoS Pathog.* 14:e1006881. doi: 10.1371/journal.ppat.1006881

Ikeda, H., Laigret, F., Martin, M. A., and Repaske, R. (1985). Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. *J. Virol.* 55, 768–777.

Imakawa, K., and Nakagawa, S. (2017). The phylogeny of placental evolution through dynamic integrations of retrotransposons. *Prog. Mol. Biol. Transl. Sci.* 145, 89–109. doi: 10.1016/bs.pmbts.2016.12.004

Iranzo, J., Cuesta, J. A., Manrubia, S., Katsnelson, M. I., and Koonin, E. V. (2017). Disentangling the effects of selection and loss bias on gene dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 114, E5616–E5624. doi: 10.1073/pnas.1704925114

Ito, J., Sugimoto, R., Nakaoka, H., Yamada, S., Kimura, T., Hayano, T., et al. (2017). Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLoS Genet.* 13:e1006883. doi: 10.1371/journal.pgen.1006883

Ito, J., Watanabe, S., Hiratsuka, T., Kuse, K., Odahara, Y., Ochi, H., et al. (2013). Refrex-1, a soluble restriction factor against feline endogenous and exogenous retroviruses. *J. Virol.* 87, 12029–12040. doi: 10.1128/JVI.01267-13

Iwasaki, Y. W., Siomi, M. C., and Siomi, H. (2015). PIWI-interacting RNA: its biogenesis and functions. *Annu. Rev. Biochem.* 84, 405–433. doi: 10.1146/annurev-biochem-060614-034258

Jeltsch, A., Kröger, M., and Pingoud, A. (1995). Evidence for an evolutionary relationship among type-II restriction endonucleases. *Gene* 160, 7–16. doi: 10.1016/0378-1119(95)00181-5

Jeltsch, A., and Pingoud, A. (1996). Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.* 42, 91–96. doi: 10.1007/BF02198833

Jimenez, R. M., Polanco, J. A., and Lupták, A. (2015). Chemistry and biology of self-cleaving ribozymes. *Trends Biochem. Sci.* 40, 648–661. doi: 10.1016/j.tibs.2015.09.001

Jung, Y. T., Lyu, M. S., Buckler-White, A., and Kozak, C. A. (2002). Characterization of a polytropic murine leukemia virus proviral sequence associated with the virus resistance gene Rmcf of DBA/2 mice. *J. Virol.* 76, 8218–8224. doi: 10.1128/JVI.76.16.8218-8224.2002

Kapitonov, V. V., and Koonin, E. V. (2015). Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. *Biol. Direct* 10:20. doi: 10.1186/s13062-015-0055-8

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Katzourakis, A., Tristem, M., Pybus, O. G., and Gifford, R. J. (2007). Discovery and analysis of the first endogenous lentivirus. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6261–6265. doi: 10.1073/pnas.0700471104

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Khan, F., Furuta, Y., Kawai, M., Kaminska, K. H., Ishikawa, K., Bujnicki, J. M., et al. (2010). A putative mobile genetic element carrying a novel type IIF restriction-modification system (PluTI). *Nucleic Acids Res.* 38, 3019–3030. doi: 10.1093/nar/gkp1221

Kobayashi, Y., Horie, M., Nakano, A., Murata, K., Itou, T., and Suzuki, Y. (2016). Exaptation of bornavirus-like nucleoprotein elements in afrotherians. *PLoS Pathog.* 12:e1005785. doi: 10.1371/journal.ppat.1005785

Koonin, E. V. (2018). Hunting for treasure chests in microbial defense Islands. *Mol. Cell* 70, 761–762. doi: 10.1016/j.molcel.2018.05.025

Koonin, E. V., and Dolja, V. V. (2014). Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol. Mol. Biol. Rev.* 78, 278–303. doi: 10.1128/MMBR.00049-13

Koonin, E. V., and Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* 16, 184–192. doi: 10.1038/nrg3859

Koonin, E. V., and Krupovic, M. (2018). The depths of virus exaptation. *Curr. Opin. Virol.* 31, 1–8. doi: 10.1016/j.coviro.2018.07.011

Koonin, E. V., and Makarova, K. S. (2017). Mobile genetic elements and evolution of CRISPR-Cas systems: all the way there and back. *Genome Biol. Evol.* 9, 2812–2825. doi: 10.1093/gbe/evx192

Kovalskaya, N., and Hammond, R. W. (2014). Molecular biology of viroid-host interactions and disease control strategies. *Plant Sci.* 228, 48–60. doi: 10.1016/j.plantsci.2014.05.006

Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D., and Koonin, E. V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol.* 12:36. doi: 10.1186/1741-7007-12-36

Lavialle, C., Cornelis, G., Dupressoir, A., Esnault, C., Heidmann, O., Vernochet, C., et al. (2013). Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368:20120507. doi: 10.1098/rstb.2012.0507

Lee, E. J., Banerjee, S., Zhou, H., Jammalamadaka, A., Arcila, M., Manjunath, B. S., et al. (2011). Identification of piRNAs in the central nervous system. *RNA* 17, 1090–1099. doi: 10.1261/rna.2565011

Li, Y., Basavappa, M., Lu, J., Dong, S., Cronkite, D. A., Prior, J. T., et al. (2016). Induction and suppression of antiviral RNA interference by influenza A virus in mammalian cells. *Nat. Microbiol.* 2:16250. doi: 10.1038/nmicrobiol.2016.250

Lima, W. F., Wu, H., Nichols, J. G., Sun, H., Murray, H. M., and Crooke, S. T. (2009). Binding and cleavage specificities of human Argonaute2. *J. Biol. Chem.* 284, 26017–26028. doi: 10.1074/jbc.M109.010835

Löber, U., Hobbs, M., Dayaram, A., Tsangaras, K., Jones, K., Alquezar-Planas, D. E., et al. (2018). Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proc. Natl. Acad. Sci. U.S.A.* 115, 8609–8614. doi: 10.1073/pnas.1807598115

Ma, B. G., Chen, L., Ji, H. F., Chen, Z. H., Yang, F. R., Wang, L., et al. (2008). Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* 366, 607–611. doi: 10.1016/j.bbrc.2007.12.014

Mack, M., Bender, K., and Schneider, P. M. (2004). Detection of retroviral antisense transcripts and promoter activity of the HERV-K(C4) insertion in the MHC class III region. *Immunogenetics* 56, 321–332. doi: 10.1007/s00251-004-0705-y

Maillard, P. V., Ciaudo, C., Marchais, A., Li, Y., Jay, F., Ding, S. W., et al. (2013). Antiviral RNA interference in mammalian cells. *Science* 342, 235–238. doi: 10.1126/science.1241930

Maillard, P. V., Van der Veen, A. G., Deddouche-Grass, S., Rogers, N. C., Merits, A., Reis, E., et al. (2016). Inactivation of the type I interferon pathway reveals long double-stranded RNA-mediated RNA interference in mammalian cells. *EMBO J.* 35, 2505–2518. doi: 10.15252/embj.201695086

Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., et al. (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* 42, 4160–4179. doi: 10.1093/nar/gkt1414

Makarova, K. S., Wolf, Y. I., Snir, S., and Koonin, E. V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.* 193, 6039–6056. doi: 10.1128/JB.05535-11

Makarova, K. S., Zhang, F., and Koonin, E. V. (2017). SnapShot: class 2 CRISPR-Cas systems. *Cell* 168, 328–328.e1. doi: 10.1016/j.cell.2016.12.038

Malfavon-Borja, R., and Feschotte, C. (2015). Fighting fire with fire: endogenous retrovirus envelopes as restriction factors. *J. Virol.* 89, 4047–4050. doi: 10.1128/JVI.03653-14

Matskevich, A. A., and Moelling, K. (2007). Dicer is involved in protection against influenza A virus infection. *J. Gen. Virol.* 88, 2627–2635. doi: 10.1099/vir.0.83103-0

McKinney, H. H. (1929). Mosaic diseases in the Canary Islands, West Africa and Gibraltar. *J. Agric. Res.* 39, 557–578.

Moelling, K., Abels, S., Jendis, J., Matskevich, A., and Heinrich, J. (2006a). Silencing of HIV by hairpin-loop-structured DNA oligonucleotide. *FEBS Lett.* 580, 3545–3550.

Moelling, K., Matskevich, A., and Jung, J. S. (2006b). Relationship between retroviral replication and RNA interference machineries. *Cold Spring Harb. Symp. Quant. Biol.* 71, 365–368.

Moelling, K., and Broecker, F. (2015). The reverse transcriptase-RNase H: from viruses to antiviral defense. *Ann. N. Y. Acad. Sci.* 1341, 126–135. doi: 10.1111/nyas.12668

Moelling, K., Broecker, F., Russo, G., and Sunagawa, S. (2017). RNase H as gene modifier, driver of evolution and antiviral defense. *Front. Microbiol.* 8:1745. doi: 10.3389/fmicb.2017.01745

Mölling, K., Bolognesi, D. P., Bauer, H., Büsen, W., Plassmann, H. W., and Hausen, P. (1971). Association of viral reverse transcriptase with an enzyme degrading the RNA moiety of RNA-DNA hybrids. *Nat. New Biol.* 234, 240–243. doi: 10.1038/newbio234240a0

Monde, K., Terasawa, H., Nakano, Y., Soheilian, F., Nagashima, K., Maeda, Y., et al. (2017). Molecular mechanisms by which HERV-K Gag interferes with HIV-1 Gag assembly and particle infectivity. *Retrovirology* 14:27. doi: 10.1186/s12977-017-0351-8

Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., and van Sinderen, D. (2013). Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* 79, 7547–7555. doi: 10.1128/AEM.02229-13

Naderer, M., Brust, J. R., Knowle, D., and Blumenthal, R. M. (2002). Mobility of a restriction-modification system revealed by its genetic contexts in three hosts. *J. Bacteriol.* 184, 2411–2419. doi: 10.1128/JB.184.9.2411-2419.2002

Nagamalleswari, E., Rao, S., Vasu, K., and Nagaraja, V. (2017). Restriction endonuclease triggered bacterial apoptosis as a mechanism for long time survival. *Nucleic Acids Res.* 45, 8423–8434. doi: 10.1093/nar/gkx576

Nandi, S., Chandramohan, D., Fioriti, L., Melnick, A. M., Hébert, J. M., Mason, C. E., et al. (2016). Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. *Proc. Natl. Acad. Sci. U.S.A.* 133, 12697–12702. doi: 10.1073/pnas.1609287113

Nuovo, G. J., Becker, J., Simsir, A., Margiotta, M., Khalife, G., and Shevchuk, M. (1994). HIV-1 nucleic acids localize to the spermatogonia and their progeny. A study by polymerase chain reaction in situ hybridization. *Am. J. Pathol.* 144, 1142–1148.

Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., et al. (2017). DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* 3, 90–98. doi: 10.1038/s41564-017-0051-0

Oliveira, N. M., Satija, H., Kouwenhoven, I. A., and Eiden, M. V. (2007). Changes in viral protein function that accompany retroviral endogenization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 17506–17511. doi: 10.1073/pnas.0704313104

Palmarini, M., Mura, M., and Spencer, T. E. (2004). Endogenous betaretroviruses of sheep: teaching new lessons in retroviral interference and adaptation. *J. Gen. Virol.* 85, 1–13. doi: 10.1099/vir.0.19547-0

Parrish, N. F., Fujino, K., Shiromoto, Y., Iwasaki, Y. W., Ha, H., Xing, J., et al. (2015). piRNAs derived from ancient viral processed pseudogenes as transgenerational sequence-specific immune memory in mammals. *RNA* 21, 1691–1703. doi: 10.1261/rna.052092.115

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Planz, O., and Stitz, L. (1999). Borna disease virus nucleoprotein (p40) is a major target for CD8^{+}-T-cell-mediated immune response. *J. Virol.* 73, 1715–1718.

Qiu, Y., Xu, Y., Zhang, Y., Zhou, H., Deng, Y. Q., Li, X. F., et al. (2017). Human virus-derived small RNAs can confer antiviral immunity in mammals. *Immunity* 46, 992–1004. doi: 10.1016/j.immuni.2017.05.006

Schneider, P. M., Witzel-Schlömp, K., Rittner, C., and Zhang, L. (2001). The endogenous retroviral insertion in the human complement C4 gene modulates the expression of homologous genes by antisense inhibition. *Immunogenetics* 53, 1–9. doi: 10.1007/s002510000288

Seed, K. D., Lazinski, D. W., Calderwood, S. B., and Camilli, A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494, 489–491. doi: 10.1038/nature11927

Shabalina, S. A., and Koonin, E. V. (2008). Origins and evolution of eukaryotic RNA interference. *Trends Ecol. Evol.* 23, 578–587. doi: 10.1016/j.tree.2008.06.005

Smit, A. F. A., Hubley, R., and Green, P. (2013–2015). *RepeatMasker Open-4.0.* Available at: http://www.repeatmasker.org

Sofuku, K., Parrish, N. F., Honda, T., and Tomonaga, K. (2015). Transcription profiling demonstrates epigenetic control of non-retroviral RNA virus-derived elements in the human genome. *Cell Rep.* 12, 1548–1554. doi: 10.1016/j.celrep.2015.08.007

Song, J. J., Smith, S. K., Hannon, G. J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* 305, 1434–1437. doi: 10.1126/science.1102514

Spencer, T. E., Mura, M., Gray, C. A., Griebel, P. J., and Palmarini, M. (2003). Receptor usage and fetal expression of ovine endogenous betaretroviruses: implications for coevolution of endogenous and exogenous retroviruses. *J. Virol.* 77, 749–753. doi: 10.1128/JVI.77.1.749-753.2003

Stetson, D. B., Ko, J. S., Heidmann, T., and Medzhitov, R. (2008). Trex1 prevents cell-intrinsic initiation of autoimmunity. *Cell* 134, 587–598. doi: 10.1016/j.cell.2008.06.032

Stoye, J. P. (2006). Koala retrovirus: a genome invasion in real time. *Genome Biol.* 7:241. doi: 10.1186/gb-2006-7-11-241

Suzuki, S. (1975). FV-4: a new gene affecting the splenomegaly induction by Friend leukemia virus. *Jpn. J. Exp. Med.* 45, 473–478.

Suzuki, Y., Frangeul, L., Dickson, L. B., Blanc, H., Verdier, Y., Vinh, J., et al. (2017). Uncovering the repertoire of endogenous flaviviral elements in aedes mosquito genomes. *J. Virol.* 91:e00571-17. doi: 10.1128/JVI.00571-17

Suzuki, Y., Kobayashi, Y., Horie, M., and Tomonaga, K. (2014). Origin of an endogenous bornavirus-like nucleoprotein element in thirteen-lined ground squirrels. *Genes Genet. Syst.* 89, 143–148. doi: 10.1266/ggs.89.143

Swarts, D. C., Makarova, K., Wang, Y., Nakanishi, K., Ketting, R. F., Koonin, E. V., et al. (2014). The evolutionary journey of Argonaute proteins. *Nat. Struct. Mol. Biol.* 21, 743–753. doi: 10.1038/nsmb.2879

Takahashi, N., Ohashi, S., Sadykov, M. R., Mizutani-Ui, Y., and Kobayashi, I. (2011). IS-linked movement of a restriction-modification system. *PLoS One* 6:e16554. doi: 10.1371/journal.pone.0016554

Tarlinton, R. E., Meers, J., and Young, P. R. (2006). Retroviral invasion of the koala genome. *Nature* 442, 79–81. doi: 10.1038/nature04841

Tarlinton, R. E., Sarker, N., Fabijan, J., Dottorini, T., Woolford, L., Meers, J., et al. (2017). Differential and defective expression of Koala Retrovirus reveal complexity of host and virus evolution. *bioRxiv* Available at: https://www.biorxiv.org/content/early/2017/11/09/211466 [accessed July 28, 2018].

tenOever, B. R. (2017). Questioning antiviral RNAi in mammals. *Nat. Microbiol.* 2:17052. doi: 10.1038/nmicrobiol.2017.52

Terry, S. N., Manganaro, L., Cuesta-Dominguez, A., Brinzevich, D., Simon, V., and Mulder, L. C. F. (2017). Expression of HERV-K108 envelope interferes with HIV-1 production. *Virology* 509, 52–59. doi: 10.1016/j.virol.2017.06.004

Tsai, K., Courtney, D. G., Kennedy, E. M., and Cullen, B. R. (2018). Influenza A virus-derived siRNAs increase in the absence of NS1 yet fail to inhibit virus replication. *RNA* 24, 1172–1182. doi: 10.1261/rna.066332.118

Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., et al. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534, 102–105. doi: 10.1038/nature17951

Wang, M., Boca, S. M., Kalelkar, R., Mittenthal, J. E., and Caetano-Anollés, G. (2006). A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. *Complexity* 12, 27–40. doi: 10.1002/cplx.20141

Wang, X., Kim, Y., Ma, Q., Hong, S. H., Pokusaeva, K., Sturino, J. M., et al. (2010). Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* 1:147. doi: 10.1038/ncomms1146

Waugh, C., Gillett, A., Polkinghorne, A., and Timms, P. (2016). Serum antibody response to koala retrovirus antigens varies in free-ranging koalas (*Phascolarctos cinereus*) in Australia: implications for vaccine design. *J. Wildl. Dis.* 52, 422–425. doi: 10.7589/2015-09-257

Weinberger, A. D., Sun, C. L., Pluciñski, M. M., Denef, V. J., Thomas, B. C., Horvath, P., et al. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput. Biol.* 8:e1002475. doi: 10.1371/journal.pcbi.1002475

Weiss, R. A. (2006). The discovery of endogenous retroviruses. *Retrovirology* 3:67.

Yang, Y., Chung, E. K., Zhou, B., Blanchong, C. A., Yu, C. Y., Füst, G., et al. (2003). Diversity in intrinsic strengths of the human complement system: serum C4 protein concentrations correlate with C4 gene size and polygenic variations, hemolytic activities, and body mass index. *J. Immunol.* 171, 2734–2745. doi: 10.4049/jimmunol.171.5.2734

Yang, Y. G., Lindahl, T., and Barnes, D. E. (2007). Trex1 exonuclease degrades ssDNA to prevent chronic checkpoint activation and autoimmune disease. *Cell* 131, 873–886. doi: 10.1016/j.cell.2007.10.017

Yuan, Y. R., Pei, Y., Ma, J. B., Kuryavyi, V., Zhadina, M., Meister, G., et al. (2005). Crystal structure of *A. aeolicus* argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. *Mol. Cell* 19, 405–419. doi: 10.1016/j.molcel.2005.07.011

Conflict of Interest Statement: 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.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 05 February 2019
doi: 10.3389/fmicb.2019.00134

Reannotation of the Ribonucleotide Reductase in a Cyanophage Reveals Life History Strategies Within the Virioplankton

*Amelia O. Harrison[1], Ryan M. Moore[2], Shawn W. Polson[2] and K. Eric Wommack[1]**

[1] School of Marine Science and Policy, University of Delaware, Newark, DE, United States, [2] Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, United States

Ribonucleotide reductases (RNRs) are ancient enzymes that catalyze the reduction of ribonucleotides to deoxyribonucleotides. They are required for virtually all cellular life and are prominent within viral genomes. RNRs share a common ancestor and must generate a protein radical for direct ribonucleotide reduction. The mechanisms by which RNRs produce radicals are diverse and divide RNRs into three major classes and several subclasses. The diversity of radical generation methods means that cellular organisms and viruses typically contain the RNR best-suited to the environmental conditions surrounding DNA replication. However, such diversity has also fostered high rates of RNR misannotation within subject sequence databases. These misannotations have resulted in incorrect translative presumptions of RNR biochemistry and have diminished the utility of this marker gene for ecological studies of viruses. We discovered a misannotation of the RNR gene within the *Prochlorococcus* phage P-SSP7 genome, which caused a chain of misannotations within commonly observed RNR genes from marine virioplankton communities. These RNRs are found in marine cyanopodo- and cyanosiphoviruses and are currently misannotated as Class II RNRs, which are O_2-independent and require cofactor B_{12}. In fact, these cyanoviral RNRs are Class I enzymes that are O_2-dependent and may require a di-metal cofactor made of Fe, Mn, or a combination of the two metals. The discovery of an overlooked Class I β subunit in the P-SSP7 genome, together with phylogenetic analysis of the α and β subunits confirms that the RNR from P-SSP7 is a Class I RNR. Phylogenetic and conserved residue analyses also suggest that the P-SSP7 RNR may constitute a novel Class I subclass. The reannotation of the RNR clade represented by P-SSP7 means that most lytic cyanophage contain Class I RNRs, while their hosts, B_{12}-producing *Synechococcus* and *Prochlorococcus*, contain Class II RNRs. By using a Class I RNR, cyanophage avoid a dependence on host-produced B_{12}, a more effective strategy for a lytic virus. The discovery of a novel RNR β subunit within cyanopodoviruses also implies that some unknown viral genes may be familiar cellular genes that are too divergent for homology-based annotation methods to identify.

Keywords: cyanophage, ribonucleotide reductase, marker gene, misannotation, cyanobacteria, viral ecology, phylogenetic analysis, virome

OPEN ACCESS

Edited by:
Jean-Michel Claverie,
Aix-Marseille Université, France

Reviewed by:
Marla Trindade,
University of the Western Cape, South Africa
Rui Zhang,
Xiamen University, China

****Correspondence:***
K. Eric Wommack
wommack@dbi.udel.edu;
wommack@udel.edu

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

Received: *06 November 2018*
Accepted: *21 January 2019*
Published: *05 February 2019*

Citation:
Harrison AO, Moore RM, Polson SW and Wommack KE (2019) Reannotation of the Ribonucleotide Reductase in a Cyanophage Reveals Life History Strategies Within the Virioplankton. Front. Microbiol. 10:134. doi: 10.3389/fmicb.2019.00134

INTRODUCTION

Viruses are the most abundant biological entities on the planet, with an estimated 10^{31} viral particles globally (Suttle, 2005). While viruses are known to infect cellular life from all three domains, viruses largely influence ecosystems through the infection of microbial hosts. In the oceans, 10^{23} viral infections are estimated to take place every second, resulting in the mortality of approximately 20% of marine microbial biomass each day (Suttle, 2007). Cell lysis resulting from viral infection influences ocean biogeochemical cycling by returning particulate and dissolved organic matter to the water column (Suttle, 2005; Jover et al., 2014), where it may be taken up by microbial populations to fuel new growth, or exported to the deep ocean (Suttle, 2007; Laber et al., 2018). Viral predation can also influence biogeochemical cycles through the restructuring of microbial populations (Rastelli et al., 2017), metabolic reprogramming of host cells (Lindell et al., 2005; Puxty et al., 2016), and horizontal gene transfer (Lindell et al., 2004).

While the importance of viruses within marine microbial communities is now commonly accepted, the biological and ecological details of viral–host interactions that influence the transformations of nutrient elements in ecosystems are largely unknown. Bridging the gap between genetic observations and ecosystem-level effects requires an understanding of the connections between genes and phenotypes. Among viruses infecting marine microbes, genes involved in nucleotide metabolism and viral replication are highly predictive of viral phenotype and evolutionary history (Iranzo et al., 2016; Kazlauskas et al., 2016; Dolja and Koonin, 2018). For example, a point mutation in motif B of the family A DNA polymerase gene (*polA*) is indicative of viral life style (Schmidt et al., 2014; Chopyk et al., 2018).

Another useful viral marker gene is ribonucleotide reductase (RNR). RNRs catalyze the rate-limiting step of DNA synthesis (ribonucleotide reduction) (Kolberg et al., 2004; Ahmad et al., 2012), and are therefore prominent in the genomes of lytic dsDNA phage (Dwivedi et al., 2013; Sakowski et al., 2014; Iranzo et al., 2016). Because RNRs have evolved into several types with diverse biochemical mechanisms and nutrient requirements (Nordlund and Reichard, 2006), the RNR used by a cell or virus can reflect the environmental conditions surrounding DNA replication (Reichard, 1993; Cotruvo et al., 2011; Sakowski et al., 2014).

All RNRs share a common catalytic mechanism in which a thiyl radical in the active site removes a hydrogen atom from the 3′ hydroxyl group of the ribose sugar, thereby activating the substrate (Licht et al., 1996; Logan et al., 1999; Lundin et al., 2015). The mechanism by which the thiyl radical is generated varies greatly among RNRs and provides the biochemical basis dividing the three major RNR classes (Lundin et al., 2015). Extant RNRs are also commonly divided by their reactivity with O_2 (Reichard, 1993): Class I RNRs are O_2-dependent; Class II RNRs are O_2-independent; and Class III RNRs are O_2-sensitive (**Figure 1A**).

Class III RNRs are the most dissimilar of the extant types, bearing no sequence similarity to Class I and II RNRs despite a common ancestry (Aravind et al., 2000; Lundin et al., 2015). They consist of two subunits that create radicals by cleaving *S*-adenosylmethionine molecules using iron-sulfur clusters (Mulliez et al., 1993). Class III RNRs are inactivated by O_2 (Eliasson et al., 1992; King and Reichard, 1995), and are therefore found only in strict or facultative anaerobes and their viruses (Fontecave et al., 2002). Class II RNRs are the only RNRs that do not require separate subunits for radical generation and catalysis (Nordlund and Reichard, 2006), and are instead encoded by a single gene, *nrdJ*. Class II RNRs require adenosylcobalamin (AdoCbl), a form of B_{12}, to produce a radical (Blakley and Barker, 1964; Lundin et al., 2010). There are two types of Class II RNR: monomeric and dimeric (Nordlund and Reichard, 2006).

Class I RNRs are the most recent (Lundin et al., 2015) and the most complex of the extant RNRs (**Figure 1B**). Radical generation takes place on a smaller subunit (β or R2) and is transferred to a larger catalytic subunit (α or R1) (Jordan and Reichard, 1998). The α subunit is encoded by *nrdA* or *nrdE* and the β subunit is encoded by *nrdB* or *nrdF*. These genes form exclusive pairs: *nrdA* is found only with *nrdB* (*nrdAB*), and *nrdE* is found only with *nrdF* (*nrdEF*). Notably, the Class I α subunit is thought to have evolved directly from dimeric Class II RNRs, so they share several catalytic sites, though sequence similarity between the two classes remains low otherwise (Lundin et al., 2015). The radical initiation mechanism of the β subunit further divides Class I RNRs into five subclasses (a–e) (Cotruvo et al., 2011, 2013; Blaesi et al., 2018; Rose et al., 2018; Srinivas et al., 2018). The subclasses are divided based on the identity of the metallocofactor (or absence thereof), the identity of the oxidant, and whether the β subunit contains (and utilizes) the tyrosine radical site (**Figure 1B**). Class I RNRs are generally presumed to be subclass Ia enzymes unless they can be assigned to another subclass based on sequence homology to a close relative that has been biochemically characterized (Berggren et al., 2017).

While the diversity of RNR biochemistry makes this enzyme an excellent marker for inferring aspects of viral biology, proper annotation of RNR genes is imperative for this purpose. Unfortunately, this same diversity has also fostered high misannotation rates, with one study reporting that 77% of RNRs submitted to GenBank had misannotations (Lundin et al., 2009). Most of those misannotations (88%) were due to RNR sequences being assigned to the wrong class. In response, a specialty database (RNRdb) was created for maintaining a collection of correctly annotated RNRs (Lundin et al., 2009). Even with resources such as the RNRdb, however, the complexity of RNR annotation remains daunting for non-experts. Class I RNRs can be particularly difficult to identify, as their classification relies largely on the annotation of both an α and β subunit.

Our prior work examining the phylogenetic relationships among RNRs from marine virioplankton revealed two large clades of cyanophage RNRs, the first made up of Class I enzymes and the second of Class II RNRs (Sakowski et al., 2014). The hosts of these cyanophage, marine *Synechococcus* and *Prochlorococcus*, carry Class II RNRs. Thus, the presence of such a large cyanophage clade with Class I RNRs was intriguing, and in contradiction to earlier findings that phage tend to carry an RNR

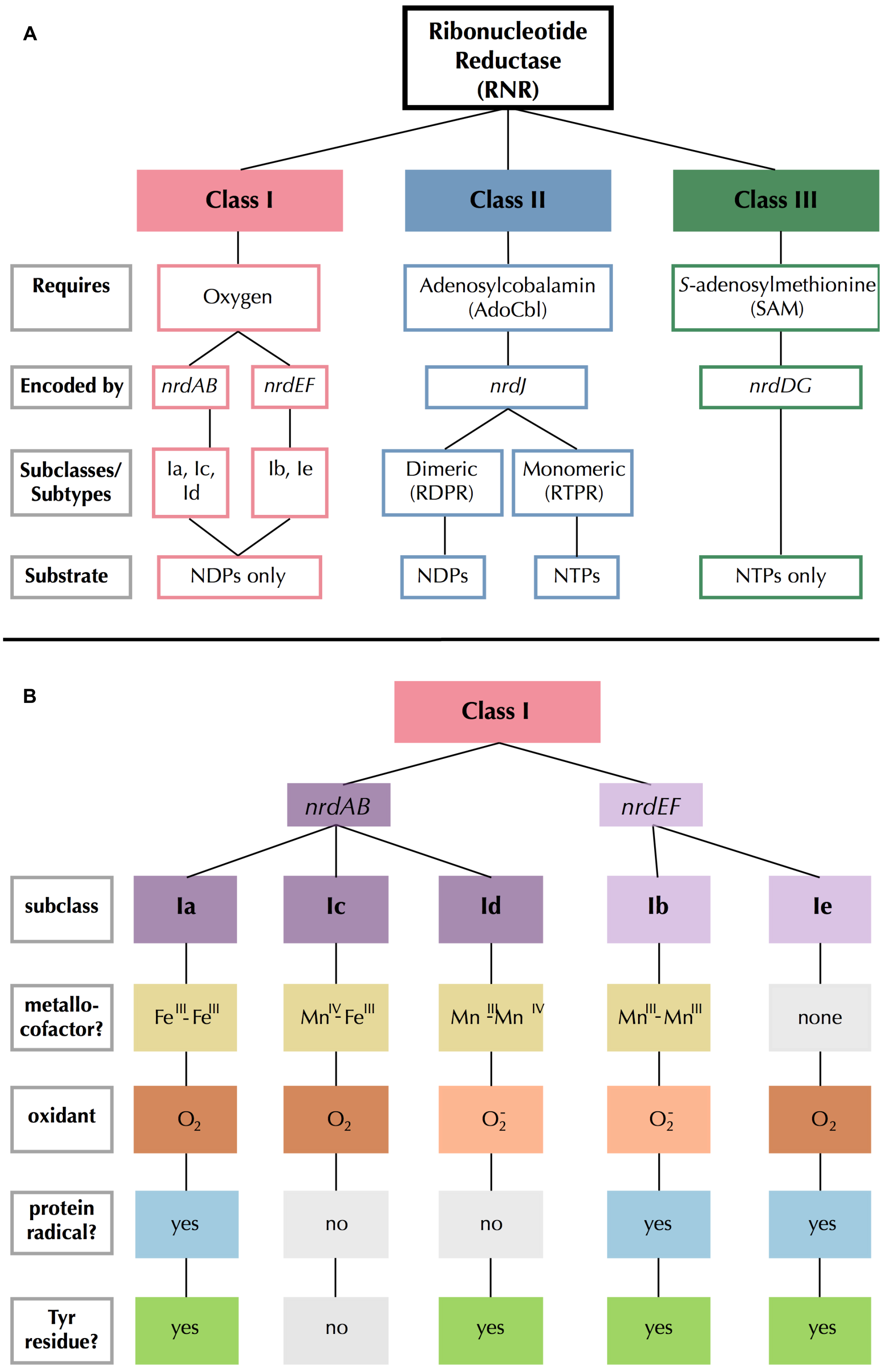


FIGURE 1 | Summary of **(A)** RNR class and **(B)** Class I subclass divisions. Gray outlined boxes to the left indicate categories. In **(B)**, like colors indicate common traits and light gray filled boxes indicate missing traits.

gene similar to that of their host cell (Dwivedi et al., 2013). Now, the reanalysis of an RNR from the Class II-carrying cyanophage has revealed that the RNRs in this second clade are, in fact, Class I RNRs that were misannotated as Class II. The reannotation of the RNR from *Prochlorococcus* phage P-SSP7 from Class II to Class I implies that most known cyanophage carry RNRs that are not host-derived, nor dependent on B_{12}. Additionally, our analysis suggests that the P-SSP7 RNR may represent a novel Class I RNR subclass.

MATERIALS AND METHODS

The Cyano SP Clade

The RNR from *Prochlorococcus* phage P-SSP7 is a member of the 'Cyano II' RNR clade, as recognized by Sakowski et al. (2014) in a study of virioplankton RNRs. Based on our analysis, and to avoid confusion with the nomenclature for RNR classes, we have renamed the Cyano II clade to the Cyano SP clade, as RNRs in this clade are exclusively found within the cyanosipho- and cyanopodoviruses (Sakowski et al., 2014). We have also renamed the Cyano I clade to the Cyano M clade, as RNRs in this clade are exclusively seen in cyanomyoviruses. The aforementioned study included ten reference sequences from the (now) Cyano SP clade. Eight of those 10 references were used in the current study (**Table 1**). Cyanophage KBS-S-1A was excluded because its genome has not been fully sequenced and *Synechococcus* phage S-CBP3 was excluded because its RNR was missing a conserved catalytic site. P-SSP7 was chosen as the clade representative because it is the most well-studied phage from this group, has a full genome available, and is the source of the original RNR misannotation.

Putative α and β Subunit Identification

Putative α and β subunit sequences were extracted from the genome of *Prochlorococcus* phage P-SSP7 (genome accession no. NC_006882.2). The putative Class I α subunit is the RNR currently identified in the P-SSP7 genome as ribonucleotide reductase class II (accession no. YP_214197.1) and was downloaded from NCBI in April 2018. As P-SSP7 has no annotated β subunit, candidate β sequences were identified based on length filtering of unannotated protein sequences. While Class I β subunits are typically between 350 and 400 amino acids (Kolberg et al., 2004), we expanded our search range to avoid excluding any potential Class I β subunits. Four candidate, unannotated proteins between 200 and 500 amino acids in length were downloaded for analysis in May 2018. Candidate proteins were searched against the Conserved Domain Database using batch CD-Search (Marchler-Bauer et al., 2017).

TABLE 1 | Cyano SP clade reference sequences and their hosts.

Virus	Family	Host	Host RNR type
Prochlorococcus phage P-SSP7	*Podoviridae*	*Prochlorococcus marinus* subsp. pastoris str. CCMP1986	II – monomeric
Cyanophage P-SSP2	*Podoviridae*	*P. marinus* MIT 9312	II – monomeric
Cyanophage 9515-10a	*Podoviridae*	*P. marinus* MIT 9515	II – monomeric
Cyanophage NATL1A-7	*Podoviridae*	*P. marinus* str. NATL1A-7	II – monomeric
Cyanophage NATL2A-133	*Podoviridae*	*P. marinus* str. NATL2A-133	II – monomeric
Cyanophage SS120-1	*Podoviridae*	*P. marinus* SS120	II – monomeric
Cyanophage Syn5	*Podoviridae*	*Synechococcus* str. WH8109	II – monomeric
Synechococcus phage S-CBS4	*Siphoviridae*	*Synechococcus* CB0101	II – monomeric

The P-SSP7 putative Class I RNR α subunit and four candidate β subunit proteins were imported into Geneious Support[1] to analyze conserved residues. The putative α subunit peptide sequence was aligned with one representative of each of the known Class I subclasses (**Supplementary Table 1**) using the MAFFT v7.388 Geneious plug-in (Katoh and Standley, 2013) on the FFT-NS-ix1000 (iterative refinement method with 1,000 iterations) setting with the BLOSUM62 scoring matrix. If necessary, alignments were manually modified to ensure that annotated active sites in the subclass representatives were properly aligned. References have been biochemically characterized and have corresponding crystal structures, where possible. Active sites were annotated for each of the subclass representatives based on literature reports and crystal structures. Residues from the putative P-SSP7 Class I α subunit aligning with active sites in subclass representatives were recorded (**Supplementary Table 2**). Candidate Class I β subunit proteins were analyzed individually in the same manner, using the β subunits corresponding to the Class I α subclass representatives (**Supplementary Table 1**). P-SSP7 candidate β subunit proteins lacking key residues were removed from the analysis. This left a single candidate β subunit protein (accession no. YP_214198.1). Putative active sites identified in the putative β subunit are recorded in **Supplementary Table 2**.

Phylogenetic Analysis

Phylogenetic Reference Sequence Curation

Sequences from the RNRdb were used as phylogenetic references. The RNRdb pulls RNRs from databases including RefSeq (Coordinators, 2014; O'Leary et al., 2016) and GenBank (Clark et al., 2016), and includes RNRs from cultured and isolated organisms and viruses as well as from environmental metagenomic samples. To create a reference sequence set for phylogenetic analyses, all available Class I α (NrdA and NrdE), Class I β (NrdB and NrdF), and Class II (NrdJ) sequences were downloaded from the RNRdb on August 20, 2018 (Lundin et al., 2009). Sequences were separated into three sets (Class I α, Class I β, and Class II) before sequence curation. Exact and sub-string matches were removed from each set using CD-HIT v4.6 (Li and Godzik, 2006; Fu et al., 2012). Sequences were then divided into smaller groups of similar sequences identified by the RNRdb. RNRdb group assignment is based on phylogenetic clade membership (Berggren et al., 2017; Rozman Grinberg et al., 2018a), so division increased sequence alignment quality. Group names and subclass membership are presented in **Table 2**. RNRdb sequences were aligned individually by group using the MAFFT v7.388 Geneious plug-in on AUTO setting

[1] https://support.geneious.com/hc/en-us/articles/227534768-How-do-I-cite-Geneious-in-a-paper-

TABLE 2 | RNR Class I subclass membership of RNRdb groups.

Class I subclass	RNRdb groups
Ia	NrdABe, NrdABg
Ia (presumed)*	NrdABh, NrdABk, NrdAm, NrdABn, NrdAq, some NrdABz (NrdABza)
Ib	some NrdEF (NrdEFb)
Ic	some NrdABz (NrdABzc)
Id	NrdABi
Ie	some NrdEF (NrdEFe)

**The Ia (presumed) subclass includes groups with no biochemically characterized members.*

with the BLOSUM62 scoring matrix. Sequence alignments were visualized and edited in Geneious v10.2.4. Inteins within RNRdb sequences were removed manually after the initial alignment step because they are evolutionarily mobile and confound phylogenetic analyses (Perler et al., 1997; Gogarten et al., 2002). After intein removal, sequences were realigned and those lacking essential catalytic residues were removed, as they are likely non-functional (Sakowski et al., 2014). Other than the two tyrosine residues involved in Class I radical transport (Y730 and Y731, *E. coli*), the same conserved residues were used for Class I α and Class II sequences (**Supplementary Table 2**). Both intein removal and catalytic residue identification for all groups were done with guidance from the annotated Class I subclass and Class II representatives (**Supplementary Table 1**).

Sequence Preparation

Broadly, three categories of phylogenies were constructed from protein sequences: (i) Class I α-only, (ii) Class I β-only, and (iii) Class I α with Class II. All phylogenies included Cyano SP clade members (**Table 1**). Class I α and Class II proteins share a common ancestor (Lundin et al., 2015), but are phylogenetically unrelated to Class I β proteins. Class I α and Class II proteins also share a common catalytic mechanism and several active sites, but are divergent enough that full-length protein sequences from both classes cannot be presented on the same phylogeny (Lundin et al., 2010). Thus, Class I α and Class II protein sequences in this analysis were trimmed to a previously defined region of interest that excluded regions not shared between the two groups (N437-S625, *E. coli* CQR81730.1) (Sakowski et al., 2014). The Class I α-only phylogeny allowed for greater resolution, as the phylogeny could be based on a longer protein sequence segment, being trimmed only before C225 in *E. coli* (CQR81730.1). Class I β sequences were trimmed to the region between W48 and Y356 (*E. coli*, KXG99827.1). For Class I α-only and Class I β-only phylogenies, sequences were trimmed near the N-terminus to exclude evolutionarily mobile ATP cone domains (Aravind et al., 2000). Class I β sequences were also trimmed near the C-terminus to exclude any fused glutaredoxin domains (Rozman Grinberg et al., 2018b). In all cases, trimming was guided by annotated Class I subclass (a-e) and Class II subtype (mono- or dimeric) representatives (**Supplementary Table 1**).

In addition to trimming, sequences were clustered prior to phylogenetic analysis, as each group contained a large number of sequences (Class I α: 15,894 sequences, Class I β: 17,109 sequences, and Class II: 9,147 sequences). To avoid inter-group mixing within individual sequence clusters, sequences were clustered by RNRdb group (**Table 2**). Clustering of RNRdb sequences was performed at multiple identity thresholds (70%, 75%, and 80%) using CD-HIT v4.7 to ensure that the placement of the Cyano SP clade was not an artifact of the identity threshold, as Cyano SP members have grouped with Class II sequences in the past (Sakowski et al., 2014). Cyano SP sequences were not clustered before phylogenetic analysis. For Class I α-only and β-only phylogenies, sequences were clustered over 80% of the alignment length. For the Class I α with Class II phylogeny, sequences were clustered over 100% of the alignment length due to the short length of the trimmed region.

Two RNRdb groups, NrdABz and NrdEF, contained member sequences belonging to two Class I subclasses (**Table 2**). In these cases, the Class I β sequences (NrdBz and NrdF) were assigned to subclasses based on active sites. For NrdBz, Class I β subunit enzymes were classified as subclass Ia (NrdBza) by the presence of a Tyr residue in the Tyr radical site (Tyr122 in *E. coli* R2), or as subclass Ic (NrdBzc) by the presence of a Phe, Leu, or Val mutation in the Tyr radical site (Lundin et al., 2009). For NrdF, Class I β subunit enzymes were classified as subclass Ib (NrdFb) or Ie (NrdFe) if carboxylate residues were conserved or missing, respectively, from the second, fourth, and fifth metal-binding sites in relation to the subclass Ib representative (**Supplementary Table 1**). Class I α sequences (NrdAz and NrdE), which could not be assigned to subclasses based on primary sequence alone, were assigned to a subclass based on the assignment of their corresponding β subunits. Class I α subunit sequences that were not able to be paired with a β subunit, or that were paired with more than one β subunit, were excluded from further analysis. Excluded Class I α subunit sequences included 1,006 NrdAz and 2,921 NrdE sequences, or 31% and 45% of total curated NrdAz and NrdE sequences, respectively. The excluded sequences comprised a small percentage of overall RNR diversity (**Supplementary Table 3**). Thus, their exclusion is not expected to have affected the phylogenetic analyses (**Supplementary Table 3**). All other RNRdb groups exclusively belonged to a single subclass.

Phylogenetic Tree Construction

For all phylogenetic analyses and clustering identity thresholds, cluster representatives were aligned with correspondingly trimmed α or β subunits from the Cyano SP clade. All alignments were constructed in Geneious using the MAFFT v7.388 plug-in with setting FFT-NS-2 (fast, progressive method) and the BLOSUM62 scoring matrix. Trees were constructed using the FastTree v2.1.5 (Price et al., 2010) Geneious plug-in with default settings. Trees were visualized and customized in Iroki (Moore et al., 2018). Phylogenies inferred from sequences clustered at different identity thresholds can be found in the **Supplementary Figures 1–3**.

Finally, a phylogeny was constructed from trimmed Class I α subunit and Class II sequences from only cyanobacteria and cyanophage. No clustering was performed. The phylogeny was constructed as described above from an alignment done using the MAFFT v7.388 plug-in with setting FFT-NS-ix1000 (iterative refinement method with 1,000 iterations).

Sequence Similarity Network

A protein sequence similarity network (SSN) was constructed with the same RNR Class I β subunit sequences used for phylogenetic analysis. The SSN was generated with the Enzyme Similarity Tool (EFI-EST) (Gerlt et al., 2015) as in Rose et al. (E-value: 5, fraction: 1, minimum alignment score: 90) (Rose et al., 2018). As the full network was too large to visualize in Cytoscape (Shannon et al., 2003; Smoot et al., 2011), the 90% identity representative node network was used (i.e., each node in the network contained sequences that shared at least 90% amino acid identity).

RESULTS

Prochlorococcus phage P-SSP7 is a cyanopodovirus that infects the marine cyanobacterium *Prochlorococcus marinus* subsp. *pastoris* str. CCMP1986 (Sullivan et al., 2005). The RNR discovered in P-SSP7 was initially annotated as Class II based on the apparent lack of a Class I β subunit in the phage genome. The RNR from P-SSP7 also lacks an ATP cone region, a domain that is common in Class I α subunits but rare in Class II enzymes (Aravind et al., 2000; Jonna et al., 2015). This was also the first cyanophage RNR of its kind to be annotated, and consequently this gene became the baseline annotation for closely related RNRs. Prior examination of RNRs in viral shotgun metagenomes (viromes) designated the phylogenetic clade containing the RNR from P-SSP7 as the 'Cyano II' clade, recognizing that member RNRs (**Table 1**), exclusively from cyanophage, were annotated as Class II and seemed to fall on the Class II side of the tree (Sakowski et al., 2014). This study also recognized a 'Cyano I' clade composed exclusively of cyanomyoviruses that carried Class I RNRs (Sakowski et al., 2014). The Cyano II clade has been renamed to Cyano SP, as the clade is comprised solely of RNRs from cyanosipho- and cyanopodoviruses. The Cyano I clade has been renamed to Cyano M, as it consists of RNRs strictly from cyanomyoviruses.

P-SSP7 Class I α Subunit Identification

The first indication that the RNR from P-SSP7 was misannotated as a Class II RNR came from the observation of two consecutive tyrosine residues (Y730 and Y731 in *E. coli*) that are present in the C-terminus of Class I α subunits and participate in long-range radical transport between the α and β subunits of Class I RNRs (Uhlin and Eklund, 1994; Greene et al., 2017). These tyrosines are not present in Class II RNRs but are present in the P-SSP7 RNR peptide (**Supplementary Table 2**). To confirm the classification of the P-SSP7 RNR as a Class I enzyme, a phylogenetic tree was constructed containing Class I α subunits and Class II sequences from the RNRdb, together with the putative α subunits from the Cyano SP clade (formerly Cyano II) reported in Sakowski et al. (2014) (**Figure 2**). Trees were constructed at different clustering identities to ensure that the placement of Cyano SP sequences with a given RNR class was not an artifact of the clustering threshold (**Supplementary Figure 1**). The Cyano SP RNRs grouped with the Class I α subunit sequences in the phylogenies constructed from sequences clustered at 75% and

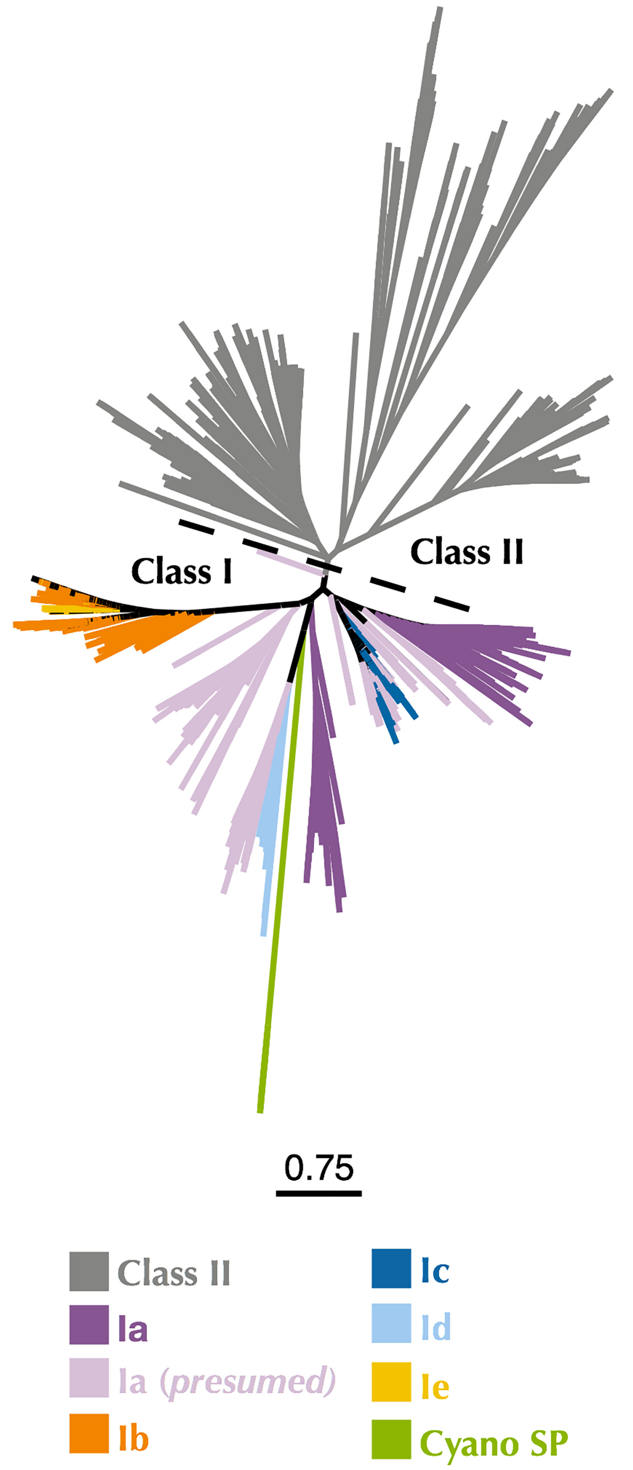


FIGURE 2 | Maximum-likelihood phylogenetic tree of Cyano SP clade α subunits with 80% clustered Class I α and Class II RNRdb sequences trimmed to a region of interest. Gray branches belong to Class II. Colored branches belong to one of the five Class I subclasses, or Cyano SP as indicated in the key. Light purple branches indicate RNRdb groups without characterized members, which are assumed to be subclass Ia enzymes. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.

80% identity, but clustered with Class II sequences in the tree made from sequences clustered at 70% identity.

P-SSP7 Class I β Subunit Identification

While the tyrosine residues within the P-SSP7 RNR are indicative of a Class I RNR, the initial annotation of the P-SSP7 RNR was made primarily because no β subunit gene could be identified within the P-SSP7 genome. Class I RNRs require a β subunit for radical generation. Because the cyanobacterial host of P-SSP7 carries a Class II RNR, the phage would have to carry its own copy of the Class I β subunit gene in order for its α subunit to function. All unannotated proteins in the P-SSP7 genome approximately the length of a Class I β subunit in the P-SSP7 genome were considered RNR β subunit candidates. Four predicted proteins within the genome matched this length criteria. A batch CD-Search (Marchler-Bauer et al., 2017) of the candidate β subunit peptide sequences was unable to identify any conserved domains in any of the sequences. Thus, we aligned the candidate P-SSP7 β subunit sequences with the sequences of biochemically characterized β subunits from each of the known Class I subclasses (**Supplementary Table 1**). Only one of the candidate sequences, accession no. YP_214198.1, was found to contain residues experimentally shown to be required for β subunit function (**Supplementary Table 2**). The hypothetical protein also resided directly downstream of the α subunit, where the β subunit is typically found (Dwivedi et al., 2013). Thus, YP_214198.1 was identified as the missing P-SSP7 β subunit.

Assignment of P-SSP7 RNR to a Class I Subclass

Class I subclasses are based on the mechanism of radical generation utilized by the β subunit. Alignment with representative Class I RNR β subunit sequences found that the P-SSP7 β subunit lacked the tyrosine residue (Y122 in *E. coli* R2) on which the stable protein radical is formed in subclasses Ia, Ib, and Ie (**Figure 1B**). The lack of the tyrosine residue seemed to indicate that the P-SSP7 β subunit belonged to subclass Ic, as Ic is the only described subclass that lacks this residue completely (the residue is conserved in Id but does not harbor a radical) (Högbom et al., 2004; Blaesi et al., 2018; Rose et al., 2018; Srinivas et al., 2018). Each subclass has a unique combination of metal-binding residues and uses a different metallocofactor (or does not bind metals at all, in the case of subclass Ie) (Blaesi et al., 2018; Srinivas et al., 2018). The residues in the putative P-SSP7 β subunit aligning with the first sphere of metal-binding residues of the subclass representatives (**Table 3**) were consistent with Class I RNRs that require metallocofactors (subclasses Ia–Id) and exactly matched subclasses Ic and Id (Blaesi et al., 2018; Srinivas et al., 2018). However, when considering second sphere binding residues, the overall pattern of metal-binding residues in the P-SSP7 β subunit did not match that of any subclass representative (**Table 3**), nor of any existing RNRdb group (**Table 4**).

Known Class I subclasses are either monophyletic or contain members that are closely related (Berggren et al., 2017; Rozman Grinberg et al., 2018a). Thus, phylogenetic trees were constructed to confirm proper subclass assignment of the P-SSP7 RNR using Class I β subunit sequences from the RNRdb clustered at 70%, 75%, and 80% and β subunits from the Cyano SP clade members. In a phylogenetic analysis of the 70% identity cluster representative sequences, the P-SSP7 β subunit and Cyano SP homologs were phylogenetically distinct from known RNRs, and did not clearly join with RNRdb groups, instead branching directly off the backbone of the tree (**Figure 3**). In the phylogenetic reconstructions at 75% and 80% identity, the Cyano SP group remained distinct but branched closely with either the NrdBg group (75% identity, subclass Ia) or the NrdBh group (80% identity, subclass Ia presumed) (**Supplementary Figure 2**). Notably, the Cyano SP β subunits branched away from subclass Ic members (NrdBzc subgroup) in all phylogenies (**Supplementary Figure 2**), making it unlikely that the Cyano SP clade belongs to subclass Ic.

Because Class I subclass assignment was inconclusive based on the β subunit metal-binding residues and phylogenetic analysis, we constructed a protein sequence similarity network (SSN) using the Enzyme Similarity Tool (EFI-EST) (Gerlt et al., 2015) as per Rose et al. (2018) with the same β subunit sequences used for phylogenetic tree construction (**Figure 4**). The SSN also provided an alignment-free method for viewing connections between RNR sequences, an especially important consideration for highly divergent peptides such as the Cyano SP clade RNRs (Gerlt et al., 2015). Most sequences were members of large, distinct subgraphs with sequences exclusively from a single RNRdb group (e.g., NrdBk and NrdBg). However, some RNRdb groups were evenly spread across multiple subgraphs of similar size (e.g., NrdBh and NrdBi), likely indicating a higher level of sequence heterogeneity

TABLE 3 | Metal-binding amino acid residues in each of the β subunit references and P-SSP7.

Organism	Subclass	First sphere						Second sphere	
		1	2	3	4	5	6	7	8
E. coli	Ia	D85	E116	H119	E205	E239	H242	S115	D238
S. typhimurium	Ib	D67	E98	H101	E158	E192	H195	M97	D191
C. trachomatis	Ic	E89	E120	H123	E193	E227	H230	E119	D226
F. johnsoniae	Id	E67	E97	H100	E160	E195	H198	C96	D194
A. urinae	Ie	D85	V116	H119	P176	K210	H213	M115	D209
P-SSP7	Cyano SP	E42	E71	H74	E117	E147	H150	D70	D146

TABLE 4 | Metal-binding amino acid residues in each of the RNRdb groups and the Cyano SP clade.

Subclass	Clade	First Sphere						Second Sphere	
		1	2	3	4	5	6	7	8
Ia Ia (presumed)	NrdBe	D	E	H	E	E	H	M/I/V	D
	NrdBg	D	E	H	E	E	H	S	D
	NrdBh	D	E	H	E	E	H	E/Q	D
	NrdBk	D	E	H	E	E	H	M/R/I	D/E
	NrdBn	D	E	H	E	E	H	E	D
	NrdBza	D	E	H	E	E	H	E	D
Ib	NrdFb	D	E	H	E	E	H	M	D
Ic	NrdBzc	E	E	H	E	E	H	E	D
Id	NrdBi	E	E	H	E	E	H	C/S	D/E
Ie	NrdFe	D	Q/V	H	S/P	K	H	M	D
If	Cyano SP	E	E	H	E	E	H	D	D

RNRdb groups are based on phylogenetic clades.

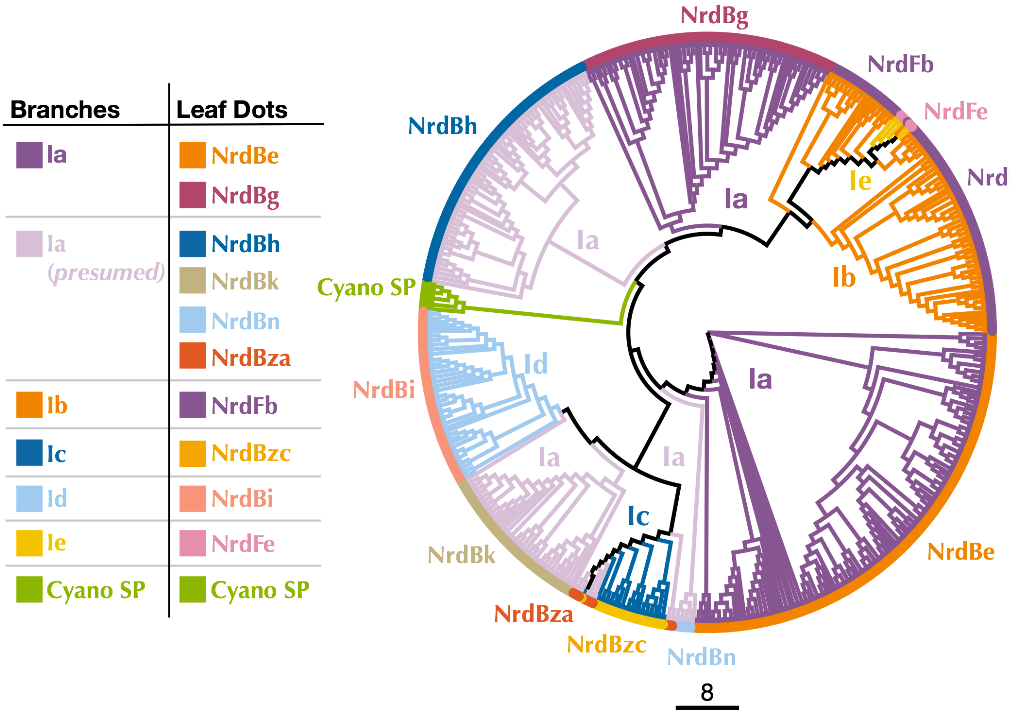


FIGURE 3 | Cladogram of near full-length Cyano SP and 70% clustered RNRdb Class I β subunit sequences. Branch colors indicate Class I subclass and leaf dot colors correspond to RNRdb group. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.

than other groups. The Cyano SP clade representatives formed exclusive subgraphs not connected to other RNRdb sequences and were divided into three singleton and one non-singleton cluster, indicating that the clade representatives are divergent even from each other.

Assignment of the Cyano SP RNRs to an existing Class I subclass could not be reliably made based on the analysis of β subunit metal-binding residues, phylogenies, or the protein SSN. Instead, the missing tyrosine radical residue, unique pattern of metal-binding sites, and phylogenetic divergence of the Cyano SP β subunits from RNRdb groups likely indicate that the Cyano SP clade represents a novel Class I subclass.

Origin of the P-SSP7 RNR

Class I α and β subunits tend to evolve in units, producing highly similar phylogenies (Lundin et al., 2010; Dwivedi et al., 2013). Because placement of the Cyano SP β subunits on phylogenetic trees changed with the percent amino acid identity used for clustering RNR sequences (**Supplementary Figure 2**), the Cyano SP α subunits were evaluated for clues to the

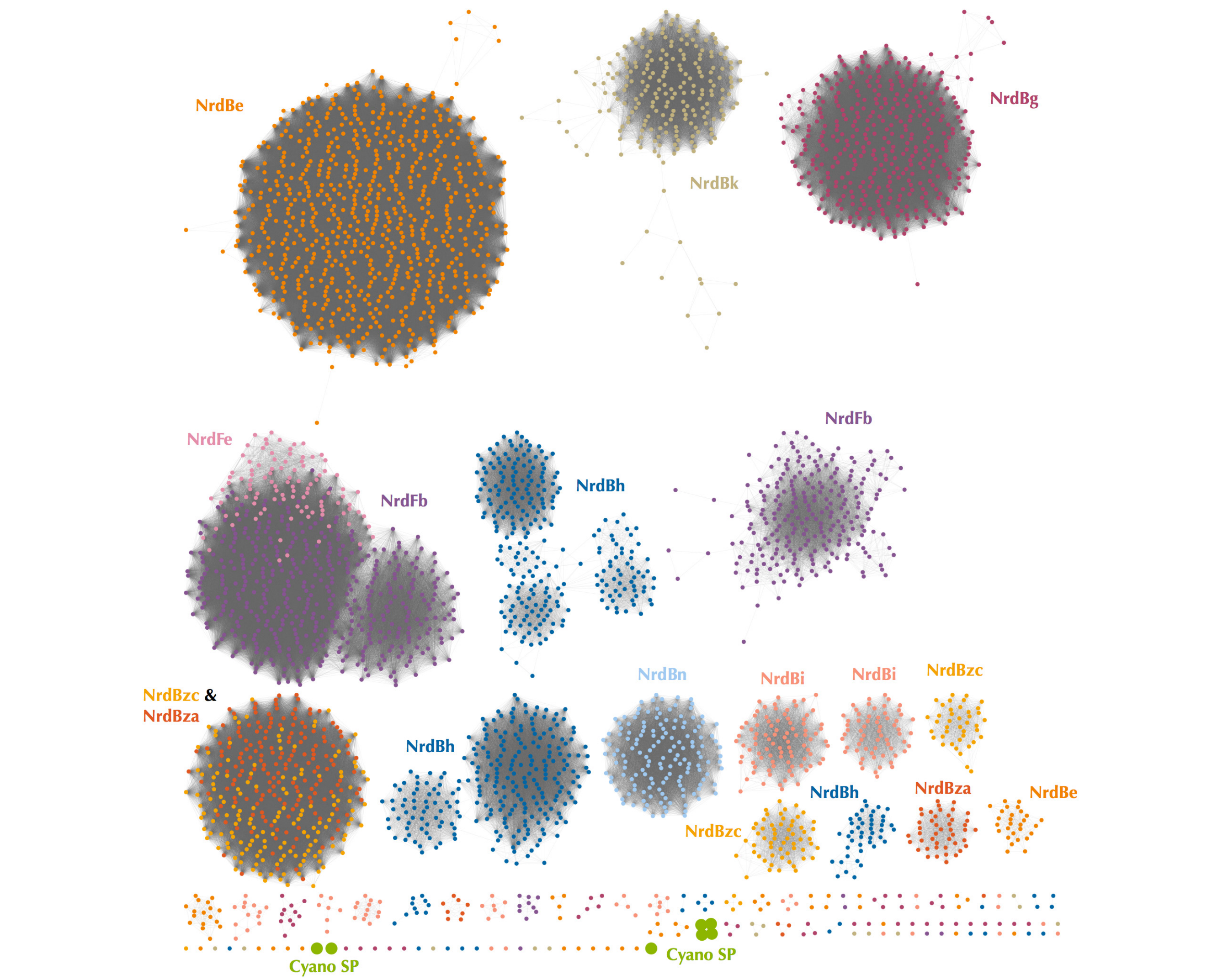


FIGURE 4 | Protein sequence similarity network of the Cyano SP clade and all RNRdb Class I β subunit sequences included in phylogenetic analysis. Nodes represent sequence clusters ≥90% similarity. Nodes are colored based on RNRdb group and match leaf dot colors on the cladogram in **Figure 3**. Edges connect nodes with minimum alignment score ≥90. Network was visualized and customized in Cytoscape.

origin of the RNR in P-SSP7. Class I α-only phylogenies were built from sequences longer than those used for the combined Class I α-Class II phylogenies, allowing greater phylogenetic resolution. Representative RNRdb Class I α subunit sequences from 70%, 75%, and 80% identity clusters were assessed. Regardless of the clustering identity, the Class I α subunit phylogenies showed consistent placement of the Cyano SP clade as an outgroup for the branch that contains RNRdb groups NrdAi (subclass Id) and NrdAk (subclass Ia presumed) (**Figure 5** and **Supplementary Figure 3**). Like the Class I β phylogenies, the Cyano SP α subunit clade was distinct and was not surrounded by any RNRdb group. The phylogenetic placement of the Cyano SP Class I α sequences among RNRdb groups (**Figure 5** and **Supplementary Figure 3**) was different from that seen for the Cyano SP Class I β sequences (**Figure 3** and **Supplementary Figure 2**). Thus, a conclusive placement for the Cyano SP β subunits among RNRdb groups was not possible.

DISCUSSION

The Cyano SP RNR Has Adapted to the Intracellular Environment

The perceived lack of a β subunit gene in the P-SSP7 genome may have led to the initial misannotation of the P-SSP7 RNR gene as a Class II RNR (Sullivan et al., 2005). Additionally, it seems unusual for a virus to carry a different class of RNR than its host (Dwivedi et al., 2013). Given that cellular organisms carry RNRs that are adapted to their environmental niche (Reichard, 1993; Cotruvo et al., 2011), viruses would also likely benefit from having the same RNR type as their host cell. The preference

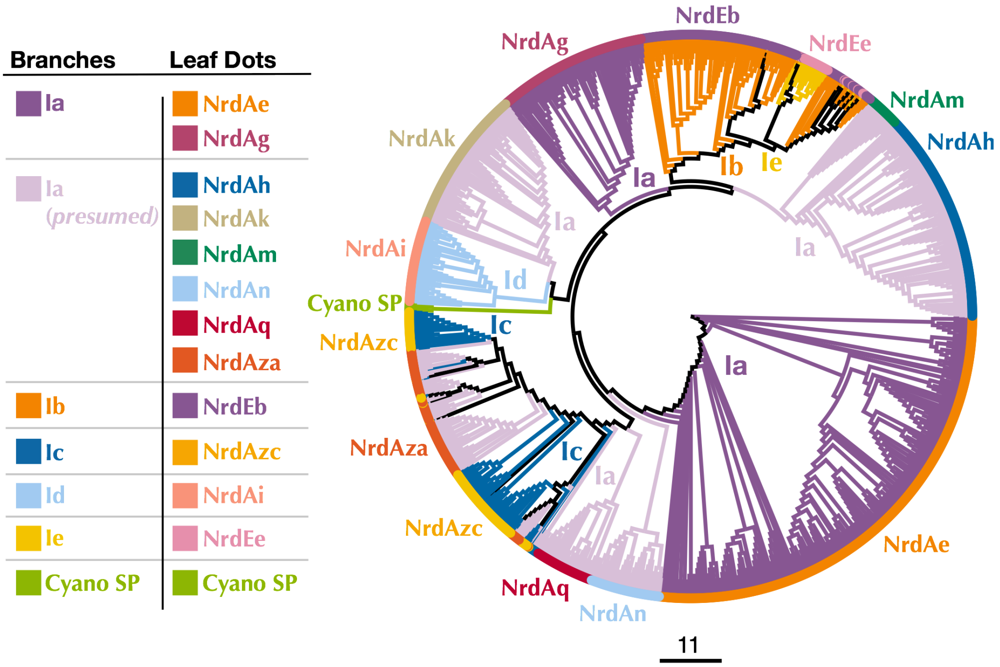


FIGURE 5 | Cladogram of near full-length Cyano SP and RNRdb Class I α subunit sequences clustered at 80%. Branch colors indicate Class I subclass and leaf dot colors correspond to RNRdb group. Colors matching to clades in **Figure 3** indicate α/β subunit pairs. Note there are α subunit clades that do not have corresponding, distinct β subunit clades, as the α subunits have diverged more than the β subunits. NrdAm β subunits belong to β subunit group NrdBh. NrdAq β subunits belong to β subunit subgroup NrdBza. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bar represents amino acid changes per 100 positions.

for a potentially iron-dependent Class I RNR enzyme among cyanophage seems puzzling considering that iron is often the primary limiting nutrient in the oceans, including in regions dominated by *Synechococcus* and *Prochlorococcus* (Moore et al., 2013; Browning et al., 2017), hosts infected by phage within the Cyano SP (cyanosipho- and cyanopodoviruses) (**Table 1**) and Cyano M (cyanomyoviruses) clades. *Synechococcus* and *Prochlorococcus* are also some of the few B_{12} producers in the oceans (Heal et al., 2016; Helliwell et al., 2016). Therefore, B_{12} availability would seem to be sufficient for viral replication with a B_{12}-dependent Class II RNR, while iron availability for phage-infected cells could be too low to support the highly lytic phenotype displayed by many of these phage.

However, carrying a Class I RNR would relieve marine cyanophage of their dependence on the host to produce sufficient levels of B_{12} for deoxyribonucleotide synthesis by a Class II enzyme. Although it is less limiting in ocean waters, B_{12} is likely to be more limiting than iron inside a cyanobacterial cell. In Cyanobacteria, B_{12} is used as a cofactor for two enzymes, the Class II RNR (NrdJ) and methionine synthase (MetH) (Heal et al., 2016). NrdJ is needed only while the cell is actively replicating, thus, transcription of this gene is closely tied with the cell cycle (Herrick and Sclavi, 2007; Mowa et al., 2009). Similarly, MetH expression is high during early growth of the B_{12}-producing cyanobacterium *Synechocystis* but decreases when cells enter the stationary growth phase (Tanioka et al., 2009). Given that NrdJ and MetH are both tied to cellular growth, intracellular B_{12} concentrations are likely highly variable.

In addition, cobalt, the metal at the center of B_{12}, is required almost exclusively for B_{12} formation and is tightly controlled because of its toxicity to cells (Waldron et al., 2009; Huertas et al., 2014). In contrast, both iron and manganese are required for numerous proteins and molecules within a cyanobacterial cell that are needed throughout the cell cycle (Palenik et al., 2003; Shcolnick and Keren, 2006). Cytoplasmic cyanobacterial iron and manganese quotas have been documented at 10^6 atoms/cell (Keren et al., 2002, 2004) and a study that aimed to identify and quantify metals in a cyanobacterium found that iron was present in high intracellular concentrations, while cobalt concentrations were below the detection limit (Barnett et al., 2012). Furthermore, some *Prochlorococcus* are able to maintain growth while up-taking just one atom of cobalt per cell per hour (Hawco and Saito, 2018). Therefore, upon infection, a cyanophage would encounter an intracellular pool of iron many fold larger than that of B_{12}.

The acquisition of B_{12} from the surrounding environment also seems unlikely. B_{12} is bulky and structurally complex, requiring special transporters which neither *Prochlorococcus*, *Synechococcus*, nor their phages are known to encode (Rodionov et al., 2003; Tang et al., 2012; Pérez et al., 2016). Furthermore, one study showed that while some organisms, such as eukaryotic microalgae, are able to import partial or finished forms of B_{12}, *Synechococcus* and likely *Prochlorococcus* are unable to do this (Helliwell et al., 2016). Instead, *Synechococcus* is required to synthesize B_{12} start to finish (Helliwell et al., 2016), likely because both *Prochlorococcus* and *Synechococcus* produce a form of B_{12} that seems to be unique to Cyanobacteria (Heal et al., 2016).

Finally, B_{12} is energetically expensive to synthesize. B_{12} synthesis requires a long pathway made up of roughly twenty different enzymes (Warren et al., 2002). By comparison, some Class I RNR metallocofactors are known to self-assemble (Cotruvo et al., 2011). At most, a metallocofactor may require a flavodoxin (NrdI) for assembly (Blaesi et al., 2018). When considering that carrying a Class I enzyme relieves the phage of relying on a complex host-mediated pathway for a molecule that is not consistently produced throughout the cell cycle, the difference in RNR type between host and phage is not surprising.

The RNR from P-SSP7 also seems to have adapted to the environment inside the host cell in other ways. The P-SSP7 β subunit lacks the tyrosine residue used for radical generation in most Class I RNR subclasses (**Figure 1B**). The tyrosine residue harbors a stable protein radical and is a target of nitric oxide (Eiserich et al., 1995; Radi, 2004). Tyrosine-radical scavenging nitric oxide is hypothesized to be present inside *Synechococcus* cells as an intermediate in nitrate reduction (Preimesberger et al., 2017), which is widespread among freshwater and marine *Synechococcus* species and is coupled to photosynthesis (Guerrero, 1985; González et al., 2006; Klotz et al., 2015; Sunda and Huntsman, 2015). Thus, the loss of the tyrosine radical site in the Class I β subunit genes of cyanophage, such as P-SSP7, would enable these phages to avoid RNR inactivation by nitric oxide.

Connections Between RNR and Cyanophage Phenotype

RNR type appears to be predictive of cyanophage morphology, as membership in a particular RNR clade corresponds to phenotype. For example, most marine cyanophage belong to the Cyano M and Cyano SP clades. The Cyano M clade consists of subclass Ia RNRs belonging solely to cyanomyoviruses. The Cyano SP clade consists of RNRs from the proposed novel subclass (If) belonging solely to cyanosipho- and cyanopodoviruses. In addition, all sequenced Class II RNRs from marine cyanophage belong exclusively to the P60 clade, which contains only cyanosipho- and cyanopodoviruses. Furthermore, the myovirus Cyanophage S-TIM5 carries a subclass Id RNR, a subclass not carried by cyanosipho- or cyanopodoviruses. Differences in the RNRs carried by different morphological groups cyanophage have been used to demonstrate possible niche exclusion in the diel infection dynamics of cyanomyovirus and cyanopodovirus populations in the Chesapeake Bay (Sakowski et al., 2014).

Most Class I RNR α subunits contain an ATP cone region. ATP cones are regulatory sites that essentially act as on/off switches for RNRs (Brown and Reichard, 1969; Aravind et al., 2000). When ATP is bound, the RNR holoenzyme enters a conformational state that allows for function (Eriksson et al., 1997). Once dNTP levels rise high enough, dATP binds the ATP cone and the holoenzyme enters a non-functional conformation (Eriksson et al., 1997; Mathews, 2006). Intriguingly, the Class I α subunits of the Cyano SP clade do not have ATP cones. This is unusual for Class I α subunits and likely represents an evolutionary loss, given that only two Class I α subunit clades (NrdAi/NrdAk and NrdEb/NrdEe) (**Figure 5**) lack ATP cones (Aravind et al., 2000; Jonna et al., 2015). In losing the ATP cone domain, the Cyano SP RNRs have lost this regulatory switch. As a consequence, the RNR of cyanopodo- and cyanosiphoviruses cannot be inactivated through dATP binding, which would be beneficial to a fast-replicating lytic phage (Chen et al., 2009).

The highly lytic nature of the Cyano SP clade is also reflected in the biochemistry of the family A DNA polymerase gene (*polA*) carried by some of the members of the clade (**Supplementary Table 4**). The amino acid residue at position 762 (*E. coli* numbering) plays a role in shaping the activity and fidelity of Pol I (polA peptide) and is hypothesized to be reflective of phage lifestyle (Schmidt et al., 2014). Prior work found that a mutation from phenylalanine to tyrosine at position 762 produced a 1,000-fold increase in processivity with a concomitant loss of fidelity (Tabor and Richardson, 1987). Three of the member phages within the Cyano SP clade carry a Pol I with a tyrosine at position 762, indicating that Cyano SP members are capable of fast DNA replication. Other members carry *polA* genes that contain a frameshift mutation, preventing identification of the 762 position. Pairing an unregulated RNR, such as the Cyano SP RNR, with a highly processive DNA polymerase would be advantageous for a highly lytic phage. This phenotype is thought to be characteristic of most cyanopodoviruses (Suttle and Chan, 1993; Wang and Chen, 2008; Schmidt et al., 2014). Observations of gene associations such as Tyr762 PolA and Cyano SP clade Class I RNR can thus inform predictions of the possible life history characteristics of unknown viruses.

A Novel Class I RNR in Cyanophage

Reannotation of the P-SSP7 RNR from Class II to Class I is based primarily on the discovery of a Class I β subunit in the P-SSP7 genome. The P-SSP7 β subunit was identified using conserved residues, as no conserved domains could be identified in the previously hypothetical protein. Our discovery of the Class I β subunit via active sites and genome location demonstrates that some unknown viral proteins (i.e., the viral genetic dark matter) (Krishnamurthy and Wang, 2017) could actually be well known proteins that are simply too divergent for annotation using homology searches or gene model approaches.

The reannotation is also supported by the presence of the consecutive tyrosine residues in the C-terminus of the newly annotated Class I α subunit, which are essential for radical transfer between Class I α and β subunits (Uhlin and Eklund, 1994; Greene et al., 2017) and are not found in Class II RNRs. Additionally, two trees constructed from Class I α and Class II sequences showed the Cyano SP clade (represented by P-SSP7) on the Class I side of the tree (**Figure 2** and **Supplementary Figure 1B**). While the 70% Class I α with Class II tree showed the Cyano SP clade on the Class II side of the tree, we believe this is an artifact of the low identity threshold and short region of interest (**Supplementary Figure 1A**). Protein SSNs constructed from the same sequences used in the Class I α with Class II phylogeny showed the Cyano SP clade as being distinct from both Class I and Class II sequences (**Supplementary Figure 4**). Thus, the high divergence of the Cyano SP clade as compared to Class I α and Class II sequences in the RNRdb are likely contributing to the Cyano SP clade grouping with Class II sequences on the 70% tree. A study of gene transcription in P-SSP7- infected *Prochlorococcus*

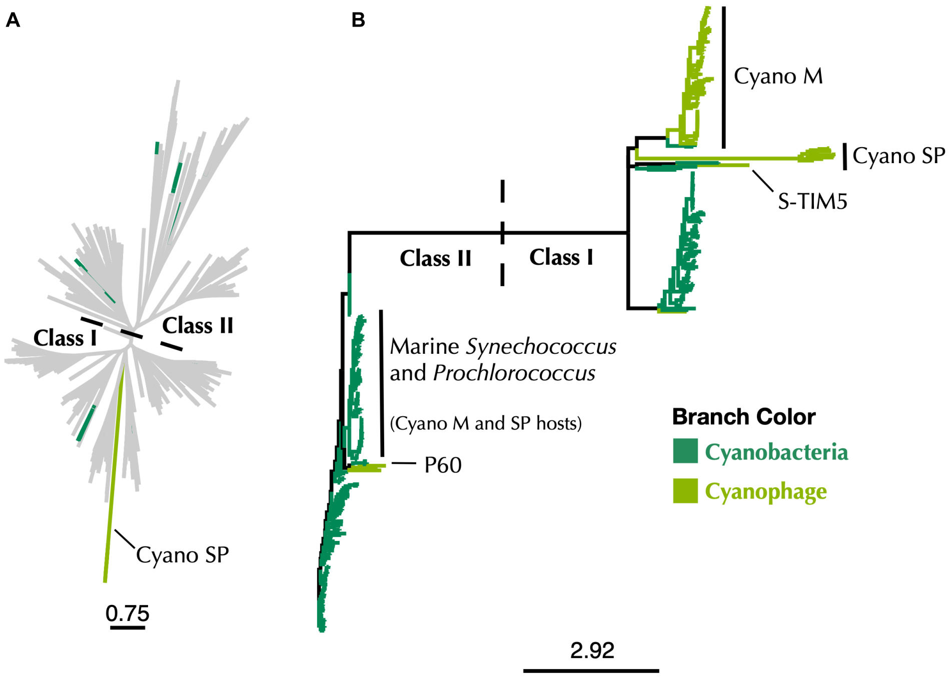


FIGURE 6 | (A) Maximum-likelihood phylogenetic tree of Cyano SP clade α subunits with 80% clustered Class I α and Class II RNRdb sequences trimmed to a region of interest. **(B)** Maximum-likelihood phylogenetic tree of a subset of Class I α subunit sequences limited to Cyanobacteria and cyanophage. In both trees, dark green branches indicate Cyanobacteria and light green branches indicate cyanophage. Trees were constructed using FastTree and visualized and customized in Iroki. Scale bars represent amino acid changes per 100 positions.

cultures lends experimental support for the presence of a Class I RNR in P-SSP7. Both the P-SSP7 Class I RNR α subunit (identified as nrd-020) and the neighboring β subunit (identified as nrd-021) were co-expressed during the second stage of phage infection, during which DNA replication typically takes place (Lindell et al., 2007).

Assignment of the P-SSP7 RNR to an existing Class I subclass was inconclusive as the radical-generating β subunit (Cotruvo et al., 2011) could not be clearly assigned based on conserved residues. β subunits are used for subclassification because, unlike α subunits that all have a consistent mechanism, the mechanisms of β subunits are variable. While the P-SSP7 β subunit contains all of the conserved residues required for function (**Supplementary Table 2**), it lacks the tyrosine residue (Y122 in *E. coli*) that harbors the stable protein radical or is conserved in subclasses Ia, Ib, Id, and Ie (Nordlund and Eklund, 1993; Cotruvo et al., 2013; Blaesi et al., 2018) (**Figure 1B**). Assignment also could not be made to subclass Ic, the only known subclass lacking the tyrosine residue (Högbom et al., 2004), based on the outcome of phylogenetic (**Figure 3** and **Supplementary Figure 2**) and protein SSN analysis (**Figure 4**). We also examined the metal-binding sites in the P-SSP7 β subunit, as metallocofactor identity is used to discriminate between subclasses Ia–Id (Cotruvo et al., 2011; Rose et al., 2018). The metal-binding residues for the P-SSP7 and other Cyano SP clade member β subunits formed a different pattern than is seen in any of the RNRdb groups (**Table 4**). The combination of the unique metal-binding residues, the lack of a tyrosine residue on which to generate a protein radical, and the phylogenetic distance between the Cyano SP clade and subclass Ic (NrdBzc) sequences, suggest that the P-SSP7 Class I β subunit may constitute a novel subclass of Class I RNRs.

Origin of the P-SSP7 RNR

Because P-SSP7's host, like most marine *Synechococcus* and *Prochlorococcus*, carries a Class II RNR, we were interested in the origin of the Class I RNR found in P-SSP7. The Class I β subunit phylogenies inconsistently placed the Cyano SP clade. Examination of Class I α subunit trees showed a consistent placement of the Cyano SP clade at the base of the branch harboring the RNRdb groups NrdAk (Ia presumed) and NrdAi (subclass Id) (**Figure 5** and **Supplementary Figure 3**). This is perhaps to be expected as, like the NrdAk and NrdAi groups, the Cyano SP Class I α subunits do not contain ATP cone domains, a trait that is rare among Class I α subunits (Jonna et al., 2015).

The observation that the Cyano SP clade does not have the same placement on the Class I β-only and Class I α-only trees is highly unusual. In viruses and cellular organisms, Class I α and β subunits are thought to evolve as units (Dwivedi et al., 2013), producing trees with the same patterns (Lundin et al., 2010). However, viral genomes are known to be highly modular, consisting of genes from multiple sources (Iranzo et al., 2016; Krupovic et al., 2018). It seems possible that an ancestral phage of the Cyano SP clade incorporated the Class I α and β subunits separately. Given that Class I α and β subunits can only perform ribonucleotide reduction as a unit, i.e., both subunits are required for functionality, these acquisitions would have had to occur in quick succession to avoid loss by the phage. Perhaps in support of this hypothesis is that the Cyano SP β subunits sometimes cluster with the NrdBg group (subclass Ia) which harbors the Cyano M clade, while the Cyano SP α subunits consistently cluster with the NrdAi group (subclass Id) that contains the *Synechococcus* phage S-TIM5. These phage groups (i.e., Cyano SP, S-TIM5, and Cyano M) all infect marine *Synechococcus* and *Prochlorococcus*, making the possibility more likely that the Cyano SP RNRs are a mosaic of these cyanomyoviral groups, with the α subunit having been acquired from a cyanophage related to S-TIM5 and the β subunit from a member of the Cyano M clade.

A phylogeny constructed using all Cyanobacteria and cyanophage present in the RNRdb with the Cyano SP clade demonstrates that the majority of known cyanophage carry Class I RNRs (**Figure 6**). The *Synechococcus* or *Prochlorococcus* hosts of phages in the Cyano M, Cyano SP clades, *Synechococcus* phage S-TIM5, and the Cyanophage P60 clade all carry Class II RNRs (Chen and Lu, 2002; Sabehi et al., 2012; Sakowski et al., 2014). Despite being a myovirus, S-TIM5 does not carry an RNR belonging to the Cyano M clade, likely because it is believed to represent a separate lineage of myoviruses (Sabehi et al., 2012). Cyanosipho- and cyanopodoviruses were found in two widely separated clades. Lytic cyanosipho- and cyanopodoviruses within the Cyanophage P60 RNR clade contain a Class II RNR, which is the same type carried by their hosts, whereas cyanosipho- and cyanopodoviruses in the Cyano SP clade contain a Class I RNR. The biological and ecological explanations behind this divergence are a mystery; however, prior work has indicated that cyanopodoviruses can be broadly divided into two clusters, MPP-A and MPP-B, based on whole genome analyses (Huang et al., 2015), but no single gene or gene group clearly distinguishes the two clusters. Nevertheless, RNRs belonging to the Cyano SP clade seem to be more common among cyanosipho- and cyanopodoviruses (Sakowski et al., 2014; Huang et al., 2015). Whether carrying a Class II RNR is the ancestral state of cyanosipho- and cyanopodoviruses could not be determined from our phylogenies.

The use of marker genes such as RNR in studying viral ecology is important in connecting genomic information to phenotypic traits. However, correct annotation of these genes is essential if accurate information is to be gained. This reannotation means that most marine cyanophage carry RNRs that did not come from their hosts (**Figure 6**), which has implications for our understanding about the acquisition of nucleotide metabolism genes by viruses. That Cyano SP clade members carry Class I RNRs and have lost the tyrosyl radical site in the β subunit is also a reminder that viruses have to adapt to the intracellular environment as well as the extracellular environment. Finally, the discovery of an overlooked β subunit implies that some unknown viral gene space may be composed of known genes that are too divergent for similarity-based annotation methods to detect but can still be identified by other means.

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the RNRdb[2]. Accession numbers for the Cyano SP clade, including genome accession, can be found in the **Supplementary Material**. The **Supplementary Material** also contains accession numbers for the annotated RNR subclass representatives.

AUTHOR CONTRIBUTIONS

AH did the analysis and wrote the manuscript. RM created the sequence similarity networks, assisted with the analysis, and edited the manuscript. KW and SP contributed to study design, data interpretation, and manuscript preparation. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Science Foundation Office of Integrated Activities, grant number 1736030 and the National Science Foundation Division of Biological Infrastructure, grant number 1356374. Computational support by the University of Delaware Center for Bioinformatics and Computational Biology Core Facility was made possible by funding from Delaware INBRE (NIH P20 GM103446) and the Delaware Biotechnology Institute.

ACKNOWLEDGMENTS

We would like to thank Barbra D. Ferrell for critical reading and input on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00134/full#supplementary-material

[2] http://rnrdb.pfitmap.org/

REFERENCES

Ahmad, M. F., Singh Kaushal, P., Wan, Q., Wijerathna, S. R., An, X., Huang, M., et al. (2012). Role of arginine 293 and glutamine 288 in communication between catalytic and allosteric sites in yeast ribonucleotide reductase. *J. Mol. Biol.* 419, 315–329. doi: 10.1016/j.jmb.2012.03.014

Aravind, L., Wolf, Y. I., and Koonin, E. V (2000). The ATP-Cone: an evolutionarily mobile, atp-binding regulatory domain. *J. Mol. Microbiol. Biotechnol.* 301, 191–194.

Barnett, J. P., Scanlan, D. J., and Blindauer, C. A. (2012). Fractionation and identification of metalloproteins from a marine cyanobacterium. *Anal. Bioanal. Chem.* 402, 3371–3377. doi: 10.1007/s00216-011-5708-6

Berggren, G., Lundin, D., and Sjöberg, B.-M. (2017). "Assembly of dimanganese and heterometallic manganese proteins," in *Encyclopedia of Inorganic and Bioinorganic Chemistry*, ed. R. A. Scott (?Hoboken, NJ: John Wiley & Sons, Ltd.). doi: 10.1002/9781119951438.eibc2480

Blaesi, E. J., Palowitch, G. M., Hu, K., Kim, A. J., Rose, H. R., Alapati, R., et al. (2018). Metal-free class Ie ribonucleotide reductase from pathogens initiates catalysis with a tyrosine-derived dihydroxyphenylalanine radical. *Proc. Natl. Acad. Sci. U.S.A.* 15, 10022–10027. doi: 10.1073/pnas.1811993115

Blakley, R. L., and Barker, H. A. (1964). Cobamide stimulation of the reduction of ribotides to deoxyribotides in *Lactobacillus leichmanii*. *Biochem. Biophys. Res. Commun.* 16, 391–397. doi: 10.1016/0006-291X(64)90363-8

Brown, N. C., and Reichard, P. (1969). Role of effector binding in allosteric control of ribonucleoside diphosphate reductase. *J. Mol. Biol.* 46, 39–55. doi: 10.1016/0022-2836(69)90056-4

Browning, T. J., Achterberg, E. P., Rapp, I., Engel, A., Bertrand, E. M., Tagliabue, A., et al. (2017). Nutrient co-limitation at the boundary of an oceanic gyre. *Nature* 551, 242–246. doi: 10.1038/nature24063

Chen, F., and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage p60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* 68, 2589–2594. doi: 10.1128/AEM.68.5.2589-2594.2002

Chen, F., Wang, K., Huang, S., Cai, H., Zhao, M., Jiao, N., et al. (2009). Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environ. Microbiol.* 11, 2884–2892. doi: 10.1111/j.1462-2920.2009.02033.x

Chopyk, J., Allard, S., Nasko, D. J., Bui, A., Mongodin, E. F., and Sapkota, A. R. (2018). Agricultural freshwater pond supports diverse and dynamic bacterial and viral populations. *Front Microbiol.* 9:792. doi: 10.3389/fmicb.2018.00792

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–72. doi: 10.1093/nar/gkv1276

Coordinators, N. R. (2014). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 42, D7–D17. doi: 10.1093/nar/gkt1146

Cotruvo, J. A., Stich, T. A., Britt, R. D., Stubbe, J., and Stubbe, J. (2013). Mechanism of assembly of the dimanganese-tyrosyl radical cofactor of class Ib ribonucleotide reductase: enzymatic generation of superoxide is required for tyrosine oxidation via a Mn(III)Mn(IV) intermediate. *J. Am. Chem. Soc.* 135, 4027–4039. doi: 10.1021/ja312457t

Cotruvo, J. A., Stubbe, J., and Stubbe, J. (2011). Class I Ribonucleotide reductases: metallocofactor assembly and repair in vitro and in vivo. *Annu. Rev. Biochem.* 80, 733–767. doi: 10.1146/annurev-biochem-061408-095817

Dolja, V. V., and Koonin, E. V. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. *Virus Res.* 244, 36–52. doi: 10.1016/J.VIRUSRES.2017.10.020

Dwivedi, B., Xue, B., Lundin, D., Edwards, R. A., and Breitbart, M. (2013). A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. *BMC Evol. Biol.* 13:33. doi: 10.1186/1471-2148-13-33

Eiserich, J. P., Butler, J., van der Vliet, A., Cross, C. E., and Halliwell, B. (1995). Nitric oxide rapidly scavenges tyrosine and tryptophan radicals. *Biochem. J.* 310 (Pt 3, 745–749. doi: 10.1042/bj3100745

Eliasson, R., Pontiss, E., Fontecaves, M., Gerezq, C., Harder$, J., Jornvallll, H., et al. (1992). Characterization of components of the anaerobic ribonucleotide reductase system from *Escherichia coli*. *J. Biol. Chem.* 267, 25541–25547.

Eriksson, M., Uhlin, U., Ramaswamy, S., Ekberg, M., Regnström, K., Sjöberg, B. -M., et al. (1997). Binding of allosteric effectors to ribonucleotide reductase protein R1: reduction of active-site cysteines promotes substrate binding. *Structure* 5, 1077–1092. doi: 10.1016/S0969-2126(97)00259-1

Fontecave, M., Mulliez, E., and Logan, D. T. (2002). Deoxyribonucleotide synthesis in anaerobic microorganisms: the class III ribonucleotide reductase. *Prog. Nucleic Acid Res. Mol. Biol.* 72, 95–127. doi: 10.1016/S0079-6603(02)72068-0

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

Gerlt, J. A., Bouvier, J. T., Davidson, D. B., Imker, H. J., Sadkhin, B., Slater, D. R., et al. (2015). Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* 1854, 1019–1037. doi: 10.1016/j.bbapap.2015.04.015

Gogarten, J. P., Senejani, A. G., Zhaxybayeva, O., Olendzenski, L., and Hilario, E. (2002). Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* 56, 263–287. doi: 10.1146/annurev.micro.56.012302.160741

González, P. J., Correia, C., Moura, I., Brondino, C. D., and Moura, J. J. G. (2006). Bacterial nitrate reductases: molecular and biological aspects of nitrate reduction. *J. Inorg. Biochem.* 100, 1015–1023. doi: 10.1016/J.JINORGBIO.2005.11.024

Greene, B. L., Taguchi, A. T., Stubbe, J., and Nocera, D. G. (2017). Conformationally dynamic radical transfer within ribonucleotide reductase. *J. Am. Chem. Soc.* 139, 16657–16665. doi: 10.1021/jacs.7b08192

Guerrero, M. G. (1985). Assimilatory nitrate reduction. *Tech. Bioprod. Photosynth.* 165–172. doi: 10.1016/B978-0-08-031999-5.50023-6

Hawco, N. J., and Saito, M. A. (2018). Competitive inhibition of cobalt uptake by zinc and manganese in a pacific *Prochlorococcus* strain: insights into metal homeostasis in a streamlined oligotrophic cyanobacterium. *Limnol. Oceanogr.* 63, 2229-2249. doi: 10.1002/lno.10935

Heal, K. R., Qin, W., Ribalet, F., Bertagnolli, A. D., Coyote-Maestas, W., Hmelo, L. R., et al. (2016). Two distinct pools of B 12 analogs reveal community interdependencies in the ocean. *Proc. Natl. Acad. Sci. U.S.A.* 114, 364–369. doi: 10.1073/pnas.1608462114

Helliwell, K. E., Lawrence, A. D., Holzer, A., Scanlan, D. J., Warren, M. J., and Smith, A. G. (2016). Cyanobacteria and eukaryotic algae use different chemical variants of vitamin B 12. *Curr. Biol.* 26, 999–1008. doi: 10.1016/j.cub.2016.02.041

Herrick, J., and Sclavi, B. (2007). Ribonucleotide reductase and the regulation of DNA replication: an old story and an ancient heritage. *Mol. Microbiol.* 63, 22–34. doi: 10.1111/j.1365-2958.2006.05493.x

Högbom, M., Stenmark, P., Voevodskaya, N., McClarty, G., Gräslund, A., and Nordlund, P. (2004). The radical site in chlamydial ribonucleotide reductase defines a new R2 subclass. *Science* 305, 245–248. doi: 10.1126/science.1098419

Huang, S., Zhang, S., Jiao, N., and Chen, F. (2015). Comparative genomic and phylogenomic analyses reveal a conserved core genome shared by estuarine and oceanic cyanopodoviruses. *PLoS One* 10:e0142962. doi: 10.1371/journal.pone.0142962

Huertas, M. J., López-Maury, L., Giner-Lamia, J., Sánchez-Riego, A. M., and Florencio, F. J. (2014). Metals in cyanobacteria: analysis of the copper, nickel, cobalt and arsenic homeostasis mechanisms. *Life* 4, 865–886. doi: 10.3390/life4040865

Iranzo, J., Krupovic, M., and Koonin, E. V. (2016). The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *mBio* 7:e00978-16. doi: 10.1128/mBio.00978-16

Jonna, V. R., Crona, M., Rofougaran, R., Lundin, D., Johansson, S., Brännström, K., et al. (2015). Diversity in overall activity regulation of ribonucleotide reductase. *J. Biol. Chem.* 290, 17339–17348. doi: 10.1074/jbc.M115.649624

Jordan, A., and Reichard, P. (1998). Ribonucleotide reductases. *Annu. Rev. Biochem.* 67, 71–98. doi: 10.1146/annurev.biochem.75.103004.142443

Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W., and Weitz, J. S. (2014). The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat. Rev. Microbiol.* 12, 519–528. doi: 10.1038/nrmicro3289

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

Kazlauskas, D., Krupovic, M., and Venclovas, È. (2016). The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res.* 44, 4551–4564. doi: 10.1093/nar/gkw322

Keren, N., Aurora, R., and Pakrasi, H. B. (2004). Critical roles of bacterioferritins in iron storage and proliferation of Cyanobacteria. *Plant Physiol.* 135, 1666–1673. doi: 10.1104/pp.104.042770

Keren, N., Kidd, M. J., Penner-Hahn, J. E., and Pakrasi, H. B. (2002). A Light-dependent mechanism for massive accumulation of manganese in the photosynthetic bacterium *Synechocystis* sp. PCC 6803 †. *Biochemistry* 41, 15085–15092. doi: 10.1021/bi026892s

King, D. S., and Reichard, P. (1995). Mass spectrometric determination of the radical scission site in the anaerobic ribonucleotide reductase of *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 206, 731–735. doi: 10.1006/BBRC.1995.1103

Klotz, A., Reinhold, E., Doello, S., Forchhammer, K., Klotz, A., Reinhold, E., et al. (2015). Nitrogen starvation acclimation in synechococcus elongatus: redox-control and the role of nitrate reduction as an electron sink. *Life* 5, 888–904. doi: 10.3390/life5010888

Kolberg, M., Strand, K. R., Graff, P., and Andersson, K. K. (2004). Structure, function, and mechanism of ribonucleotide reductases. *Biochim. Biophys. Acta Proteins Proteomics* 1699, 1–34. doi: 10.1016/j.bbapap.2004.02.007

Krishnamurthy, S. R., and Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Res.* 239, 136–142. doi: 10.1016/J.VIRUSRES.2017.02.002

Krupovic, M., Cvirkaite-Krupovic, V., Iranzo, J., Prangishvili, D., and Koonin, E. V. (2018). Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res.* 244, 181–193. doi: 10.1016/J.VIRUSRES.2017.11.025

Laber, C. P., Hunter, J. E., Carvalho, F., Collins, J. R., Hunter, E. J., Schieler, B. M., et al. (2018). Coccolithovirus facilitation of carbon export in the North Atlantic. *Nat. Microbiol.* 3, 537–547. doi: 10.1038/s41564-018-0128-4

Li, W., 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

Licht, S., Gerfen, G. J., and Stubbe, J. (1996). Thiyl radicals in ribonucleotide reductases. *Science* 271, 477–481. doi: 10.1126/science.271.5248.477

Lindell, D., Jaffe, J. D., Coleman, M. L., Futschik, M. E., Axmann, I. M., Rector, T., et al. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449, 83–86. doi: 10.1038/nature06130

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M., and Chisholm, S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438, 86–89. doi: 10.1038/nature04111

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., and Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11013–11018. doi: 10.1073/pnas.0401526101

Logan, D. T., Andersson, J., Sjöberg, B. -M., and Nordlund, P. (1999). A Glycyl radical site in the crystal structure of a class III ribonucleotide reductase. *Science* 283, 1499–1504. doi: 10.1126/science.283.5407.1499

Lundin, D., Berggren, G., Logan, D. T., and Sjöberg, B. -M. (2015). The origin and evolution of ribonucleotide reduction. *Life* 5, 604–636. doi: 10.3390/life5010604

Lundin, D., Gribaldo, S., Torrents, E., Sjoberg, B. -M., and Poole, A. M. (2010). Ribonucleotide reduction – horizontal transfer of a required function spans all three domains. *BMC Evol. Biol.* 10:383. doi: 10.1186/1471-2148-10-383

Lundin, D., Torrents, E., Poole, A. M., and Sjöberg, B. -M. (2009). RNRdb, a curated database of the universal enzyme family ribonucleotide reductase, reveals a high level of misannotation in sequences deposited to Genbank. *BMC Genomics* 10:589. doi: 10.1186/1471-2164-10-589

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi: 10.1093/nar/gkw1129

Mathews, C. K. (2006). DNA precursor metabolism and genomic stability. *FASEB J.* 5, 1300–1314. doi: 10.1096/fj.06-5730rev

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., et al. (2013). Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* 6, 701–710. doi: 10.1038/ngeo1765

Moore, R. M., Harrison, A. O., McAllister, S. M., and Wommack, K. E. (2018). Iroki: automatic customization and visualization of phylogenetic trees. *bioRxiv* [Preprint]. doi: 10.1101/106138

Mowa, M. B., Warner, D. F., Kaplan, G., Kana, B. D., and Mizrahi, V. (2009). Function and regulation of class I ribonucleotide reductase-encoding genes in mycobacteria. *J. Bacteriol.* 191, 985–995. doi: 10.1128/JB.01409-08

Mulliez, E., Fontecave, M., Gaillard, J., and Reichard, P. (1993). An iron-sulfur center and a free radical in the active anaerobic ribonucleotide reductase of *Escherichia coli*. *J. Biol. Chem.* 268, 2296–2299.

Nordlund, P., and Eklund, H. (1993). Structure and function of the *Escherichia coli* ribonucleotide reductase protein R2. *J. Mol. Biol.* 232, 123–164. doi: 10.1006/jmbi.1993.1374

Nordlund, P., and Reichard, P. (2006). Ribonucleotide reductases. *Annu. Rev. Biochem.* 75, 681–706. doi: 10.1146/annurev.biochem.75.103004.142443

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

Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., et al. (2003). The genome of a motile marine *Synechococcus*. *Nature* 424, 1037–1042. doi: 10.1038/nature01943

Pérez, A. A., Rodionov, D. A., Bryant, D. A., Perez, A. A., Rodionov, D. A., Bryant, D. A., et al. (2016). Identification and regulation of genes for cobalamin transport in the cyanobacterium *Synechococcus* sp. Strain PCC 7002. *J. Bacteriol.* 198, 2753–2761. doi: 10.1128/JB.00476-16

Perler, F. B., Olsen, G. J., and Adam, E. (1997). Compilation and analysis of intein sequences. *Nucleic Acids Res.* 25, 1087–1093. doi: 10.1093/nar/25.6.1087

Preimesberger, M. R., Johnson, E. A., Nye, D. B., and Lecomte, J. T. J. (2017). Covalent attachment of the heme to *Synechococcus* hemoglobin alters its reactivity toward nitric oxide. *J. Inorg. Biochem.* 177, 171–182. doi: 10.1016/J.JINORGBIO.2017.09.018

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Puxty, R. J., Millard, A. D., Evans, D. J., and Scanlan, D. J. (2016). Viruses inhibit CO_2 fixation in the most abundant phototrophs on earth. *Curr. Biol.* 26, 1585–1589. doi: 10.1016/j.cub.2016.04.036

Radi, R. (2004). Nitric oxide, oxidants, and protein tyrosine nitration. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4003–4008. doi: 10.1073/pnas.0307446101

Rastelli, E., Corinaldesi, C., Dell'Anno, A., Tangherlini, M., Martorelli, E., Ingrassia, M., et al. (2017). High potential for temperate viruses to drive carbon cycling in chemoautotrophy-dominated shallow-water hydrothermal vents. *Environ. Microbiol.* 19, 4432–4446. doi: 10.1111/1462-2920.13890

Reichard, P. (1993). From RNA to DNA, Why so many ribonucleotide reductases? *Science* 260, 1773–1777.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2003). Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J. Biol. Chem.* 278, 41148-41159. doi: 10.1074/jbc.M305837200

Rose, H. R., Ghosh, M. K., Maggiolo, A. O., Pollock, C. J., Blaesi, E. J., Hajj, V., et al. (2018). Structural basis for superoxide activation of flavobacterium johnsoniae Class I ribonucleotide reductase and for radical initiation by its dimanganese cofactor. *Biochemistry* 57, 2679–2693. doi: 10.1021/acs.biochem.8b00247

Rozman Grinberg, I., Lundin, D., Hasan, M., Crona, M., Jonna, V. R., Loderer, C., et al. (2018a). Novel ATP-cone-driven allosteric regulation of ribonucleotide reductase via the radical-generating subunit. *eLife* 7:e31529. doi: 10.7554/eLife.31529

Rozman Grinberg, I., Lundin, D., Sahlin, M., Crona, M., Berggren, G., Hofer, A., et al. (2018b). A glutaredoxin domain fused to the radical-generating subunit of ribonucleotide reductase (RNR) functions as an efficient RNR reductant. *J. Biol. Chem.* 293, 15889–15900. doi: 10.1074/jbc.RA118.004991

Sabehi, G., Shaulov, L., Silver, D. H., Yanai, I., Harel, A., and Lindell, D. (2012). A novel lineage of myoviruses infecting cyanobacteria is widespread in the oceans. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2037-2042. doi: 10.1073/pnas.1115467109

Sakowski, E. G., Munsell, E. V, Hyatt, M., Kress, W., Williamson, S. J., Nasko, D. J., et al. (2014). Ribonucleotide reductases reveal novel viral diversity and predict biological and ecological features of unknown marine viruses. *Proc. Natl. Acad. Sci. U.S.A.* 111, 15786-15791. doi: 10.1073/pnas.1401322111

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

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

Shcolnick, S., and Keren, N. (2006). Metal homeostasis in cyanobacteria and chloroplasts. Balancing benefits and risks to the photosynthetic apparatus. *Plant Physiol.* 141, 805–810. doi: 10.1104/pp.106.079251

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. -L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431–432. doi: 10.1093/bioinformatics/btq675

Srinivas, V., Lebrette, H., Lundin, D., Kutin, Y., Sahlin, M., Lerche, M., et al. (2018). Metal-free ribonucleotide reduction powered by a DOPA radical in Mycoplasma pathogens. *Nature* 563, 416–420. doi: 10.1038/s41586-018-0653-6

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* 3:e144. doi: 10.1371/journal.pbio.0030144

Sunda, W. G., and Huntsman, S. A. (2015). High iron requirement for growth, photosynthesis, and low-light acclimation in the coastal cyanobacterium *Synechococcus bacillaris*. *Front. Microbiol.* 6:561. doi: 10.3389/fmicb.2015.00561

Suttle, C. A. (2005). Viruses in the sea. *Nature* 437, 356–361. doi: 10.1038/nature04160

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

Suttle, C. A., and Chan, A. M. (1993). Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics. *Mar. Ecol. Prog. Ser.* 92, 99–109. doi: 10.3354/meps092099

Tabor, S., and Richardson, C. C. (1987). DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 84, 4767–4771. doi: 10.1073/pnas.84.14.4767

Tang, K., Jiao, N., Liu, K., Zhang, Y., and Li, S. (2012). Distribution and functions of tonb-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. *PLoS One* 7:e41204. doi: 10.1371/journal.pone.0041204

Tanioka, Y., Yabuta, Y., Yamaji, R., Shigeoka, S., Nakano, Y., Watanabe, F., et al. (2009). Occurrence of Pseudo vitamin B 12 and its possible function as the cofactor of cobalamin-dependent methionine synthase in a cyanobacterium *Synechocystis* sp. PCC6803. *J. Nutr. Sci. Vitaminol.* 55, 518–521. doi: 10.3177/jnsv.55.518

Uhlin, U., and Eklund, H. (1994). Structure of ribonucleotide reductase protein R1. *Nature* 370, 533–539. doi: 10.1038/370533a0

Waldron, K. J., Rutherford, J. C., Ford, D., and Robinson, N. J. (2009). Metalloproteins and metal sensing. *Nature* 460, 823–830. doi: 10.1038/nature08300

Wang, K., and Chen, F. (2008). Prevalence of highly host-specific cyanophages in the estuarine environment. *Environ. Microbiol.* 10, 300–312. doi: 10.1111/j.1462-2920.2007.01452.x

Warren, M. J., Raux, E., Schubert, H. L., and Escalante-Semerena, J. C. (2002). The biosynthesis of adenosylcobalamin (vitamin B 12). *Nat. Prod. Rep.* 19, 390–412. doi: 10.1039/b108967f.

Conflict of Interest Statement: 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.

frontiers in Genetics

HYPOTHESIS AND THEORY
published: 12 February 2019
doi: 10.3389/fgene.2019.00065

Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation

Bhaskar Chandra Mohan Ramisetty and Pavithra Anantharaman Sudhakari*

Laboratory of Molecular Biology and Evolution, School of Chemical and Biotechnology, SASTRA Deemed University, Thanjavur, India

Bacterial genomes are highly plastic allowing the generation of variants through mutations and acquisition of genetic information. The fittest variants are then selected by the econiche thereby allowing the bacterial adaptation and colonization of the habitat. Larger genomes, however, may impose metabolic burden and hence bacterial genomes are optimized by the loss of frivolous genetic information. The activity of temperate bacteriophages has acute consequences on the bacterial population as well as the bacterial genome through lytic and lysogenic cycles. Lysogeny is a selective advantage as the prophage provides immunity to the lysogen against secondary phage attack. Since the non-lysogens are eliminated by the lytic phages, lysogens multiply and colonize the habitat. Nevertheless, all lysogens have an imminent risk of lytic cycle activation and cell lysis. However, a mutation in the attachment sites or in the genes that encode the specific recombinase responsible for prophage excision could result in 'grounding' of the prophage. Since the lysogens with grounded prophage are immune to respective phage infection as well as dodge the induction of lytic cycle, we hypothesize that the selection of these mutant lysogens is favored relative to their normal lysogenic counterparts. These grounded prophages offer several advantages to the bacterial genome evolution through propensity for genetic variations including inversions, deletions, and insertions via horizontal gene transfer. We propose that the grounded prophages expedite bacterial genome evolution by acting as 'genetic buffer zones' thereby increasing the frequency as well as the diversity of variations on which natural selection favors the beneficial variants. The grounded prophages are also hotspots for horizontal gene transfer wherein several ecologically significant genes such as those involved in stress tolerance, antimicrobial resistance, and novel metabolic pathways, are integrated. Moreover, the high frequency of genetic changes within prophages also allows proportionate probability for the *de novo* genesis of genetic information. Through sequence analyses of well-characterized *E. coli* prophages we exemplify various roles of grounded prophages in *E. coli* ecology and evolution. Therefore, the temperate prophages are one of the most significant drivers of bacterial genome evolution and sites of biogenesis of genetic information.

Keywords: genome plasticity, genome evolution, horizontal gene transfer, bacterial ecology, bacteriophage

OPEN ACCESS

Edited by:
Arshan Nasir,
COMSATS Institute of Information Technology, Pakistan

Reviewed by:
Sylwia Bloch,
University of Gdańsk, Poland
Jianhe Sun,
Shanghai Jiao Tong University, China
Alicja Wegrzyn,
University of Szczecin, Poland

****Correspondence:***
Bhaskar Chandra Mohan Ramisetty
ramisettybcm@biotech.sastra.edu;
bcmr25@gmail.com

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *28 August 2018*
Accepted: *24 January 2019*
Published: *12 February 2019*

Citation:
Ramisetty BCM and Sudhakari PA (2019) Bacterial 'Grounded' Prophages: Hotspots for Genetic Renovation and Innovation. Front. Genet. 10:65. doi: 10.3389/fgene.2019.00065

INTRODUCTION

Genome plasticity is an essential requirement for evolution directed by 'econiche.' Bacterial genomes are highly plastic relative to other organisms owing to the diversity and dynamicity of the niche (Dobrindt et al., 2010). The success of bacterial 'omnipresence' is predominantly dependent on the propensity and probability of genomic plasticity. The frequent exposure to and the ability to integrate exogenous genetic information enhances their genome plasticity (Hacker and Carniel, 2001; Dobrindt et al., 2010; Darmon and Leach, 2014; D'Souza et al., 2015). The basic principles of bacterial genomic plasticity comprise of (i) acquisition of exogenous genetic information and (ii) deletion of unnecessary genetic information at the population level by the selection imposed by the niche. The niche, including the inter and intraspecies competition, is the prime factor that imposes the selection of beneficial genetic information and optimization of the genomes based on the costs and benefits of each variation (**Figure 1**). The intraspecies competition achieves deletion of non-beneficial genetic information wherein the individuals with a higher metabolic burden, and slow growth rate are eliminated under natural testing conditions thereby selecting individuals with lesser 'junk' DNA (Dobrindt and Hacker, 2001). Furthermore, bacteria with deleterious recombination events or mutations are also eliminated thereby gradually 'ridding off' of frivolous genetic information from the population through a process generally referred to as 'purifying selection.' In other words, the econiche selects the best 'composites of genetic information' of the available ones based on the reproductive potential and stress tolerance imposed by the habitat (Forde et al., 2004).

Genome plasticity is predominantly mediated by horizontal gene transfer (HGT) and intragenomic recombinations such as transposition. The three principal HGT mechanisms are (i) transformation (direct uptake of DNA), (ii) transduction (DNA transfer mediated by phages), and (iii) conjugation (DNA transfer through physical contact between two bacteria) (Davison, 1999; von Wintersdorff et al., 2016). Of the three HGT mechanisms, transduction is highly 'active' by virtue of viral tropism, and mechanisms of DNA injection into the recipient bacterium. Occasionally, some of the temperate phages reversibly integrate into the host genome by a phenomenon referred to as lysogeny. The integrated phage genome, now referred to as 'prophage,' is vertically inherited by the daughter cells; the prophage replicates as a part of the host genome. Eventually, the prophage would switch from lysogenic to lytic cycle often resulting in phage multiplication and host cell death (Gandon, 2016). Bacteriophages, in general, play an important role in the ecology as well as the evolution of bacteria through several types of interactions between the host and phage genomes (Roossinck, 2011; Nasir et al., 2017). In fact, phages and their hosts have coevolved strategies that impact the survival, persistence, and evolution of their respective genomes (Buckling and Brockhurst, 2012; Koskella and Brockhurst, 2014; Nasir et al., 2017). Prophages were shown to influence the microbial community through enhanced recombinations (Nadeem and Wahl, 2017; Braga et al., 2018).

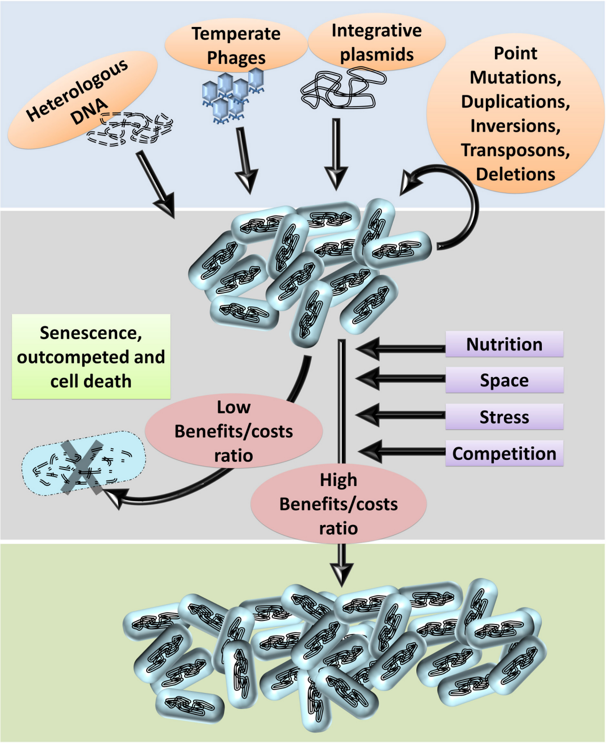

Heterologous DNA
Temperate Phages
Integrative plasmids
Point Mutations, Duplications, Inversions, Transposons, Deletions
Senescence, outcompeted and cell death
Low Benefits/costs ratio
High Benefits/costs ratio
Nutrition
Space
Stress
Competition

FIGURE 1 | Bacterial genome evolution. Based on the habitat, bacterial genomes acquire DNA (genetic information) from the environment. Genetic information from the phages, plasmids, and dead organisms maybe integrated into the genome by homologous or non-homologous recombination.

Mutations in one or more genetic elements required for the excision result in the failure of prophage excision from the host genome rendering them 'grounded,' also referred to as 'cryptic' or defective prophages (Casjens, 2003; Wang et al., 2010). The whole genome sequencing of diverse bacterial species has revealed multiple prophages within each genome. The prevalence of multiple grounded prophages within most bacterial genomes is indicative of eco-evolutionary selection of such genomes. Although the advantages of prophages are being explored (Bobay et al., 2014; Menouni et al., 2015; Gandon, 2016; Howard-Varona et al., 2017), the ecological scenarios and the possible evolutionary impact of these prophages are still unclear. In this manuscript, we discuss the mechanisms of grounding, and propose eco-evolutionary perspectives rationalizing the prevalence and advantages of grounded prophages to the host bacteria. We performed sequence analyses of five grounded prophages (DLP12, e14, Rac, CPZ-55, and Qin) of *Escherichia coli* to illustrate the distribution and the advantages.

Furthermore, several types of mutations can occur in the genome such as point mutations, duplications, inversions, deletions, etc. Within a population of bacteria, these genetic variations occur at a low frequency. Often, these variations are deleterious: i.e., the benefits caused by the variations are less than the costs incurred. Several parameters, such as nutrition, stress tolerance, competition, and resistance to antibiotics impose stringent selective pressure on the population. Hence, bacteria with deleterious variations usually perish or are outcompeted by the faster growing kin. Bacteria with beneficial genetic variations, such as higher growth rates, novel metabolic pathways, higher tolerance to stress, and higher competitive fitness fare better and are selected.

Each habitat may pose various stresses of varying degrees, allowing the gradual selection of organisms optimal for successful propagation in the prevailing conditions.

BACTERIAL GENOME PLASTICITY

Genome plasticity drives the successful spatiotemporal propagation of bacteria in terms of ecology as well as evolution. Genetic rearrangements, caused by transformation, transposition, and recombination events (site-specific as well as homologous), enhanced by insertion sequences (IS) elements, integrons, conjugative transposons, plasmids, bacteriophages, pathogenicity islands, and genomic islands result in a gain (or loss) of genes with diverse functions (Dobrindt and Hacker, 2001; Betts et al., 2018). Different bacteria may have different degrees of plasticity (Gaudriault et al., 2006) and, furthermore, regions such as 'Regions of Genomic Plasticity' (RGPs) of a bacterial genome are thought to have a higher propensity for genetic rearrangements relative to other core regions. The four different RGPs categories (genomic islands, prophages, regions encoding recombinases, and unclassified) are mosaics consisting of several modules. Each module ranging from 0.5 to 60 kb consists of genes involved in metabolism, intra and intercellular DNA mobility, drug resistance, host/environmental interactions, and antibiotic synthesis (Ogier et al., 2010). Different bacterial species have varying potential to take up exogenous DNA, metabolize it as a nutrient, or recombine the incoming DNA into the genome. The ability to take up exogenous DNA is termed as competence, a complex mechanism of temporal regulation for the uptake as well as the integration of the exogenous DNA (Blackstone and Green, 1999; Palchevskiy and Finkel, 2006; Maughan et al., 2010; Sinha and Redfield, 2012; Mell and Redfield, 2014). Often, the DNA could be catabolized and reused in nucleic acid metabolism or for generation of energy depending on the metabolic state of the cell (Finkel and Kolter, 2001). The DNA that is taken up maybe recombined into the bacterial host genome by homologous recombination or non-homologous mechanisms resulting in transformation (Fischer et al., 2001; Blokesch, 2017; Veening and Blokesch, 2017). Integration of exogenous DNA is the largest contributor of the pan-genome, the cumulative genetic information of all strains of a single bacterial species. The total amount of incoming DNA during the existence of a species in a habitat, theoretically, would imply an enormous genome. However, bacterial populations tend to get rid of larger genomes by intraspecies competition in limiting conditions imposed by the niche. The reduction of genome size is evident from the occurrence of multiple pseudogenes and loss of several genes within the genomes of most bacteria (Lawrence et al., 2001).

A typical bacterial species has expanded the horizons of its habitat through genome plasticity. When non-pathogenic bacteria infect a host, the niche may select the individual bacterium that may have specific mutations/genetic information allowing survival/virulence and eventually evolve into a pathogenic strain (Hogardt et al., 2007). Such mutations confer fitness by the acquisition of virulence genes and loss of antivirulence genes that nullify the action of factors conferring virulence (Dobrindt et al., 2010). For example, the presence of antivirulence genes encoding enzymes involved in arabinose catabolism has reduced the virulence in non-pathogenic *Burkholderia thailandensis*. This has been experimentally shown wherein *B. pseudomallei* that harbors arabinose assimilation operon, grown with arabinose as the only carbon source showed *Salmonella*-like gene cluster down-regulation (TTSS3-Type III secretion system) involved in virulence (Moore et al., 2004). *Shigella* and enteroinvasive *E. coli* (EIEC) species have acquired their virulence through structural mutation of the *cadA* gene. Another mechanism involved in pathoadaptation is the mutation of *cis*-regulatory elements (CRE). For example, mutational study in *Salmonella* species, on the role of *cis*-regulatory elements, revealed that evolutionary mutations in *cis*-regulatory elements such as SsrB regulatory system enhanced the expression of *srfN* gene involved in intrahost fitness (Osborne et al., 2009). The enormous genetic variability in *Helicobacter pylori*, a human pathogen, shows the extent of adaptation brought about by HGT and homologous recombination. However, these sequence variations have occurred without a drastic increase in the genome size of the bacterium (Kraft et al., 2006). Among closely related *Salmonella* serotypes, extensive genetic diversity was observed in the antigenic determinant genes that encode factors responsible for lipopolysaccharides (LPS), flagella, fimbriae and virulence factors (Fierer and Guiney, 2001). Genome plasticity allows oscillation of acquisition and loss of genetic information from the population as per the change in growth conditions and the habitats.

PREVALENCE OF PROPHAGES IN BACTERIAL GENOMES

It is conceivable that the prophages are the largest exogenous contributors of bacterial genetic information, primarily, because of their sheer size. Astonishingly, prophages may constitute 10–20% of a bacterial genome predominantly due to an exposure to a multitude of bacteriophages (Casjens, 2003). Of the completely sequenced *E. coli* strains, *E. coli* O157:H7 str. Sakai has the highest number of prophages per genome; it harbors 18 prophage elements constituting 16% of the total genome (Asadulghani et al., 2009). Often, these prophages could be activated by several stresses to undergo lytic cycle. For example, 40 out of 148 *Lactobacilli* strains, representing 15 species, released phage particles upon mitomycin induction. In particular, 77% of 30 *L. Salivarius* produced temperate phages, killer particles or defective phages. Temperate prophages were found in 10% of 105 strains of *L. bulgaricus* and *L. lactis* (Sechaud et al., 1988). Using VirSorter tool (Roux et al., 2015a), around 5,492 microbial genomes out of the 14,973 publicly available bacterial and archaeal genomes (as on January 2015) had viral-like elements of which 2,445 contained more than one detectable phage elements. 82% of such phage co-infections involved multiple *Caudovirales* (Roux et al., 2015b). Bacterial genomes may contain intact prophages, LT2 isolate of *Salmonella enterica* (four prophages), pathogenic *Streptococcus pyogenes* isolate SF370 (prophage SF370.1) (Canchaya et al., 2002), *H. influenza* Rd (Flu-Mu) and *E. coli* sakai (Sp18) or defective prophages like two non-inducible prophages (Lp1 and Lp2) and two prophage remnants (R-Lp3 and R-Lp4) of *Lactobacillus plantarum* (Ventura et al., 2003). A large number of IS elements that are collectively referred to as minichromosomes comprise highly divergent prophage remnants constituting 22% of *Halobacterium* sp. NRC-l genome, and code for transposases associated with phage genomes and other mobile genetic elements (Bonneau et al., 2004). A recent web based tool called PhageWeb, for the identification and characterization of prophages within bacterial genomes, may expand the exploration of prophages in a variety of species (Sousa et al., 2018). The prevalence of prophages in bacterial genomes is intriguing, firstly, because of the imminent danger of lytic cycle induction and, secondly, the metabolic burden of replicating and expressing the prophage genetic elements. The eco-evolutionary rationales that operate to sustain prophage rich genomes of bacterial populations in the environment are poorly understood. It is indeed interesting to understand the beneficiaries of lysogeny: the host bacterial genome or the phage genome or both.

WHY ARE PROPHAGES PREVALENT IN BACTERIAL GENOMES?

(a) Lysogeny Mechanisms

Lysogeny is a common feature in almost all habitats including the gut microbiome (Howard-Varona et al., 2017; Kim and Bae, 2018). Integration into the host genome is the fundamental step for the establishment of lysogeny. Different phages employ different mechanisms (**Figure 2**) that broadly fall into (i) site-specific recombination (Landy and Ross, 1977), and (ii) transposition based integration (Allet, 1979; Harshey, 2014). Site-specific recombination is a well-characterized mechanism of phage (specifically lambda bacteriophages) genome integration and is the only one discussed here for brevity. Site-specific recombination requires phage *attP* and host *attB* sites that share 15 nucleotides in common, and thus the crossover occurs either within or flanking regions of this core region (Landy and Ross, 1977). Despite defective *attB* site Int-mediated recombination occurs between *attP* and secondary attachment sites in the host genome, albeit at a lower frequency (Shimada et al., 1975). The Cre- and Int-dependent recombination between the *lox* sites are mediated by site-specific recombination system of P1 and λ phages, respectively. The *loxP* has two 17 bp binding site for Cre, and each of the sites has 13 bp inverted repeat and a 4 bp region which together forms 8 bp spacer region flanked by two 13 bp inverted repeats. The *loxP* and *loxB* sites share limited homology, and hence their recombination efficiency is altered with sequence variation. Cre recombinase makes 6 bp long double strand staggered cut in the spacer region, localized between the adjacent *loxP* sites followed by strand exchange. The process of Cre-mediated cleavage involves phosphodiester bond breakage at 3' end and covalent interaction of the primed phosphate with the protein (Hoess and Abremski, 1985). A similar rejoining mechanism is mediated by λ Int recombinase (Craig and Nash, 1983). Four recognition sites form a synapse prior to the cleavage, and this process is homology dependent (Ross et al., 1979). In essence, the essentials required for integration or excision are the *att* sites (on the bacterial as well as the phage genomes) and the specific recombinase that recognizes the concomitant recombinase recognition sites present within the corresponding *att* sites. Different recombinases, or here integrases, recognize different sequence specificities wherein the recognized sequences may have considerable sequence differences. Cross-activity of a recombinase of one phage with *att* site of another is proportional to the degree of homology between their integrases and similarity between the core- and arm-type sites in attachment sites of the respective phages (Groth and Calos, 2004).

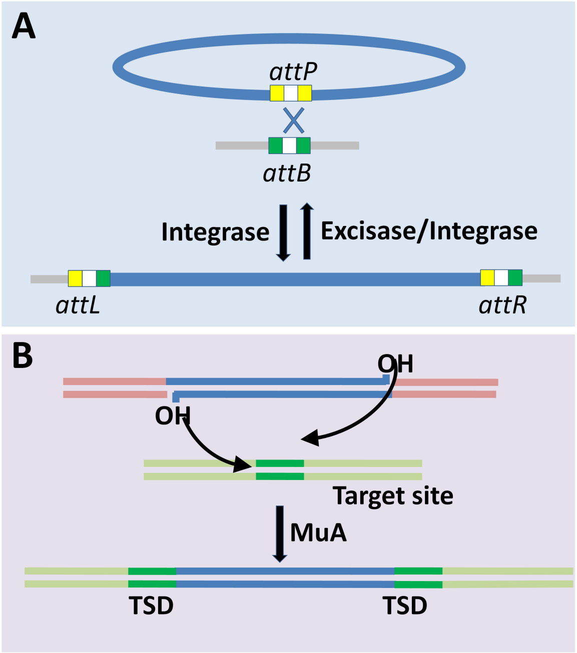


FIGURE 2 | Mechanisms of bacteriophage genome integration into bacterial chromosome. **(A)** Site-specific recombination mediated-lysogenization is well characterized for lambdoid phages. Site-specific recombination is catalyzed by the integrase through phage and bacterial attachment sites (*attP* and *attB,* respectively) resulting in recombinant *attL* and *attR* sites (Landy and Ross, 1977; Menouni et al., 2015). Prophage induction is mediated by excisase and integrase complex. **(B)** Transposition based integration of Mu phage into the target site in the host genome requires MuA transposase. The integration results in target site duplication (TSD) (Allet, 1979; Harshey, 2014).

(b) Hotspots for Prophage Integration

The fundamental premise is that any genome with deleterious recombination events, concerning the niche conditions, will be lost in the population whereas tolerable recombination events are unaffected. Occasionally, advantageous recombination events would be selected and that would eventually outcompete their kin. Integration of most prophages is by site-specific recombination at *att* sites or by transposition at a random site. Although there is a significant chance for the occurrence of the *att* sites at multiple loci of a bacterial genome, many integration events could be deleterious for the host genome because of gene disruptions or dysregulations leading to greater fitness costs. Various analyses of lysogens have shown that prophages are predominantly found at loci closer to non-coding genes (e.g., tRNA genes), a few functional genes, and intergenic regions. Hence, these integration events should be assumed as either tolerable or advantageous in the growth conditions of the context. For example, in *Ralstonia solanacearum* the *attB* site for ϕRSA1 prophage integration occurred downstream of the tRNA-Arg gene (Fujiwara et al., 2008) and in *E. coli*, the relatives of λ phage and P4 phage insert within the tRNA genes (Campbell, 2003). It is rational for a phage with a broader range of host specificity to prefer integration sites that are within protein-coding functional genes as their occurrence is conserved across distantly related species, subject to the tolerable consequences of the recombination (Hill et al., 1989). Based on the composition and length of the integration site, a variable number of sites within a host genome for the phage integration are possible. However, Mu-like prophages show random integration events due to transposition-based integration. A robust analysis of 471 *E. coli* and *Salmonella* prophages showed 58 distinctive integration loci for 369 *E. coli* prophages and 102 loci for *S. enterica* prophages (Bobay et al., 2013). Based on the empirical evidence of various prophages in numerous bacterial genomes, one may assume that integration of phage genome into that of the bacterial genome at the specified loci are either tolerable or advantageous in the ecological context in which these lysogens would be selected.

(c) Advantages of Lysogeny

The size of the prophages (approximately 45 kb each) and the expression of prophage genes impose a significant metabolic burden on the lysogen growing in stringent natural conditions competing with non-lysogenic peers. It is conceivable that the bacterial cells, which do not harbor any prophages, have a competitive growth advantage relative to lysogens. However, as noted earlier, the prevalence of multiple cryptic prophages indicate that the lysogenic strains had a survival advantage compared to the non-lysogens. Intriguingly, lysogeny, as a trait of the bacterial genome as well as phage genome, must have been selected by virtue of some spatiotemporal advantage to both the host bacterium and the phage (Paterson et al., 2010; Roossinck, 2011). One of the straight forward explanations is that some of the prophages contribute genes of selective advantages, such as antibiotic resistance conferring genes encoding multidrug resistance pump, outer membrane protease, cell division inhibiting factors, small toxic membrane polypeptide, (Wang and Wood, 2016) etc., in the habitat and hence the lysogens would outcompete the non-lysogens. However, harboring large prophages that do not confer a selective advantage is a metabolic burden and hence is a fitness disadvantage to the lysogen. The key to deciphering this paradox—selection despite the metabolic burden—is the vulnerability of non-lysogens to phage-induced lysis while the lysogens are immune to the corresponding phage. Prophages bestow lysogens with immunity against secondary phage infections, referred to as superinfection exclusion (or immunity) (**Figure 3**). Assuming that the ecological niche of the bacterial population contains the phages, the lysogens have a survival advantage due to the 'superinfection immunity' conferred by the corresponding phage. Conferring immunity is ecologically vital for the resident prophage as well as the host bacterium. Lysogeny in conjunction with superinfection immunity confers the selective advantage to the lysogen from

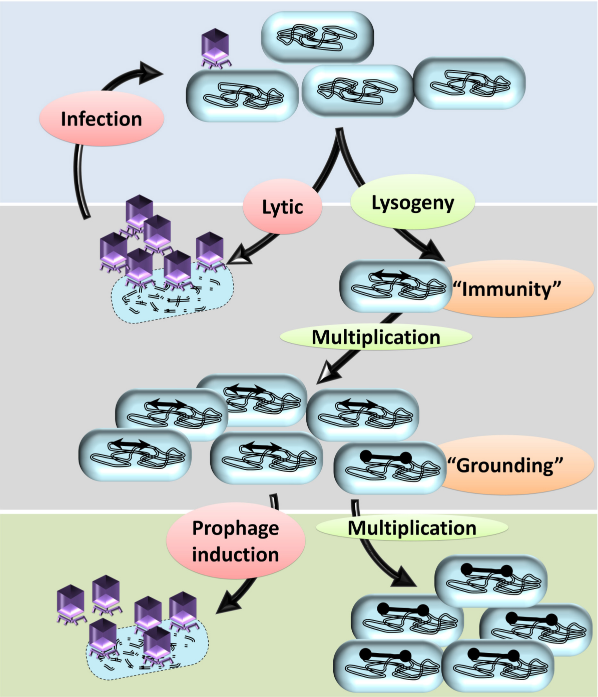


FIGURE 3 | A hypothetical model to explain the prevalence of cryptic prophages in bacterial genomes. A typical temperate phage injects its genome in the bacterial cells that may undergo either lytic cycle or lysogeny. In the lytic cycle, more copies of phages are produced, the host bacterium is lysed, and a multitude of new phages are released into the surrounding. The lytic cycle repeats with each bacterium, which almost annihilates the bacterial population. In some bacteria, the injected genome undergoes lysogeny by integrating into the bacterial genome. Lysogeny allows vertical propagation of prophage along with that of the bacterial genome. By means of the prophage, the lysogens attain immunity to the infecting secondary phages thereby preventing infection/lysis. Apart from other mutants, the lysogens are the only surviving kin and hence will multiply well due to lack of intraspecies competition. Lysogeny, upon reversal, will ultimately lead to a lytic cycle leading to the death of the host bacterium. In some of the prophages, mutations in the recombinase recognition sites or recombinase genes occur, which prevent the elicitation of the lytic cycle: a phenomenon referred to as 'grounding of prophages.' The bacteria with grounded prophages have two advantages; immunity from attack by related secondary phages and irreversible lysogeny that prevents phage-mediated cell death. Eventually, other lysogens undergo lytic cycle upon induction by various stimuli resulting in lysis of the host bacterium. The bacterium with cryptic prophages will survive and have a growth advantage due to reduced competitive pressure from kin.

phage attack and concomitantly also allows the propagation of the lysogen. Superinfection immunity, as a mechanism to prevent multiple infections of a lysogen by similar phages, is achieved by different mechanisms such as: (i) conformational change in phage receptor sites, (ii) modifications in DNA transfer system, (iii) inhibition of phage lysozyme for penetration, (iv) incoming phage DNA breakdown, and (v) abortive exclusion. The gp15 expression of HK97 had shown resistance to HK97 and its close relative HK75 infections (Cumby et al., 2012). Similar factors include (i) SieA and SieB of *Salmonella* phage P22 blocks the DNA entry with unknown component of DNA transfer system and leads to abortive infection, respectively (Susskind et al., 1974), and (ii) Sie_{2009} of *Lactococcus lactis* phage Tuc_{2009} (Mahony et al., 2008). Rex system of λ prophage is thought to induce superinfection exclusion via abortive exclusion. RexA (a component of cell cytoplasm) is activated by protein–DNA interaction which then mobilizes RexB, a membrane ion channel protein, thereby altering the membrane potential of the host (Parma et al., 1992). The CI repressor (a positive regulator of its own expression) of the incumbent λ phage inhibits the establishment of superinfecting λ phage upon forming CI repressor-DNA complex and thus preventing the integration and replication of the secondary λ phage (Gottesman and Weisberg, 2004; Fogg et al., 2010). The defective e14 prophage-encoded Lit protease upon its activation by GoI peptide (located within the major head protein) cleaves the elongation factor EF-Tu, thus blocking the late gene expression of T4 phage (McGrath et al., 2002). Upon phages attack on a bacterial population, the only survivors are the lysogens and the spontaneously evolved mutants that are no longer recognized by the phage. Such scenarios would allow the elimination of the non-lysogens resulting in survival and growth of the lysogens due to extensive relaxation of peer competition for resources. Hence, the lysogens and the mutants will have unhindered access to the resources and the 'spoils' of the dead kin allowing the repopulation of the habitat, a phenomenon similar to that of bottlenecking.

It is also proposed that the prophage genes could confer a competitive advantage at a population level such as in the case of biofilm formation (Wang et al., 2010). The individual cost of prophage induction is explained as a mechanism of altruism since the induced lysis of the lysogens that benefits their population as a whole. For example, phage-induced lysis of the lysogens in the biofilm further enhances the biofilm formation by providing additional nutrients, bacterial dispersal, and extracellular DNA. Similarly, the bacteria competing for resources produce bacteriocins and require transport protein for their release. In cases where the transport protein is not encoded, like the group B colicins producing operon in *E. coli* and *Shigella*, the phage-induced lysis promotes the release of colicin bacteriocin (Obeng et al., 2016). Shiga toxin-producing *E. coli* are lysogens for lambdoid prophages that also encode Shiga toxin. The prophage induction is followed by the expression of Shiga toxin genes release of the Shiga toxin (Iversen et al., 2015). Shiga toxin targets the eukaryotic protein synthesis of cells that can internalize the toxin. In a small subset of the lysogens the prophage is induced to lytic cycle by stimuli like H_2O_2 (Shimizu et al., 2009; Los et al., 2010, 2013). The lysis of the bacteria causes the release of Shiga toxin which acts on their predators (protozoan or neutrophils as the case maybe) (Dydecka et al., 2017; Chakraborty et al., 2018). The attenuation or inactivation of the predators is thought to increase the survivability of the rest of the population. It is hypothesized that these Shiga converting prophages is an example of altruism at the population level. However, the concept of altruism is still debatable in the case of bacteria (Ramisetty et al., 2015; Durand et al., 2016; Peeters and de Jonge, 2018). Specifically with free-living bacteria, altruistic behavior would benefit the self and/or kin if only the beneficiaries of the donation are the kin. The ultimate altruism in the form of suicide is detrimental to the unicellular free-living species unless the fraction of donors is very small. Bacteria in nature occur as multispecies colonies/biofilms wherein intraspecies-altruism is highly unlikely because the sacrifice by the donors may benefit the competing species equally. Feeding the competitors is counterproductive to the goals of altruism. In the case of prophage-mediated host lysis, the selective induction of prophage, the mechanisms limiting the induction to a small fraction of the population, and the beneficiaries of the altruistic behavior require more experimental evaluation (Los et al., 2013).

Multiple other ecological advantages of lysogeny were also reported. Phage conversion upon transduction alters the host physiology with respect to metabolism, pathogenicity, and niche adaptation (Nasir et al., 2017). The integrated auxiliary metabolic genes (AMGs) among the marine viromes have been reported to carry essential genes for different nutrient cycles such as *dsr* and *sox* genesin sulfur cycle, *amok* and P-II in nitrogen cycle, photosynthesis (*psbA* and *psbD* genes from *Prochlorococcus* and *Synechococcus* cyanophages) (Roux et al., 2016), motility proteins (e.g., *flaB*), and flagellar motor complexes (e.g., *motA*) (Hurwitz et al., 2015). Prophages increase the survival (Steinberg and Levin, 2007), cell growth (Picozzi et al., 2015), virulence (Waldor and Mekalanos, 1996; Zschöck et al., 2000), promote biofilm formation, phase variation (Kutsukake and Iino, 1980; Silverman and Simon, 1980; Sekulovic and Fortier, 2015), quorum sensing (Hargreaves et al., 2014), confer resistance to antibiotics/environmental stresses (Hyder and Streitfeld, 1978; Lopez et al., 2012; Wipf et al., 2014), and immunotolerance (Farrant et al., 1997). Yet another example for prophage-enhanced environmental stress resistance of the lysogen is phage-regulated acid resistance in *E. coli*. Comparative analysis of non-lysogen and lysogen of Shiga-toxigenic phage ($\Phi 24_B$), revealed the upregulation of operons encoding type I fimbriae and the glutamic acid decarboxylase (GAD) acid stress island (Veses-Garcia et al., 2015). It is highly likely that different prophages may have conferred one or more of the above discussed eco-evolutionary advantages. More such rationales could also be expected with the exploration and molecular examination of prophages in many other bacterial genomes. Of all the advantages discussed, immunity to superinfection is the most likely and effective advantage conferred by the prophage.

(d) 'Grounding' of Prophages

Similar to transposons, the prophages are also prone to 'grounding,' which is to imply irreversible lysogeny due to genetic events that prevent their excision. Multiple *trans* and *cis* elements are involved in the site-specific recombination of the phage with the host genome. The site-specific integration of phage (specifically lambdoid bacteriophage) into their specific host genome via host and phage attachment sites ("*attB*" and "*attP*" sites, respectively) yields the recombinant "*attL*" and "*attR*" sites. *attL* and *attR* sites are the essential *cis* elements for excision of the integrated prophage upon switching from lysogenic to lytic cycle. The prerequisites for the integration and excision of phage genomes are (i) phage-encoded recombinase proteins (integrase and excisase); (ii) conserved prophage flanking "*attR*" and "*attL*" excisase recognition sites; and (iii) the expression of accessory host factors such as IHF and factor for inversion stimulation (Fis) proteins. Efficient excision is achieved by the cooperative role of both integrase and excisase. It has been shown that Xis protein promotes excision by recognizing the phage '*attR*' site in addition to sequence-specific binding of Int or IHF protein (Yin et al., 1985). The assembly of the excisive complex and the rate of prophage curing depends on the amount of Xis expressed in the host. The host Fis enhances the binding specificity of Xis thereby promoting the formation of excisive intasome on *attR* site (Papagiannis et al., 2007). The phage cII protein, along with IHF, upregulates the expression of Int and late gene repressors and thus promotes lysogeny. The mRNA level of the integrase gene in *S. aureus* prophage is influenced by the host alternative sigma factor σ^H (Tao et al., 2010). The Bxb1 prophage of *Mycobacterium* species encodes an accessory protein, gp47, which is involved in the directionality of Bxb1 integrase for the assembly of excisive intasome (Ghosh et al., 2006). Similarly, gp3 of *streptomyces* phage ΦBT (Zhang et al., 2013), gp52 of *streptococcus* bacteriophage ΦJoe, Rv1584c from ΦRv1 phage of *Mycobacterium tuberculosis* (Bibb et al., 2005), and SprA from SPβ prophage of *Bacillus subtilis* (Abe et al., 2014) are recombination directionality factors (RDFs). The insertion of the IS element and further rearrangement in lambdoid prophage of *Shigella dysenteriae* rendered them defective upon deletion of *stx* region (McDonough and Butterton, 1999). Similarly, the RNA-mediated translational control of *int* expression by *sib* results in the formation of a truncated integrase which in turn hampers the excision of defective prophages (Schindler and Echols, 1981).

The intracellular concentration of proteins involved in recombination such as Int, IHF, Xis, and Fis is highly regulated. IHF inhibition of excisive recombination occurs at lower levels of Int protein (Thompson et al., 1986). Hence, the mostly likely events for the grounding of temperate prophages are: (i) defective recombinase proteins (Swalla et al., 2003), (ii) homologous recombination during the super infection of related phages (Canchaya et al., 2003), and (iii) resected *att* sites. DNA invertase, Xis, and Fis could promote genetic rearrangement like DNA inversion, specifically between the *attP* site and secondary attachment site within the prophage. Such genetic rearrangements could also affect the expression of phage-encoded proteins and modify the orientation of the attachment sites resulting in the establishment of irreversible lysogeny or 'grounding' (Dorgai et al., 1993).

(e) Degeneration of Prophages

The grounded prophages are highly prone to deletions, especially the genes whose products do not confer any advantage or are regressive. Harboring large prophages and expressing the prophage genes is a metabolic burden. Hence, minimizing such genetic elements are beneficial to the host genome. The closely related *E. coli* K-12 and *Shigella flexneri* showed variability in their prophage remnants (Brussow et al., 2004). A study on the genome diversity of *Salmonella enterica* serovar Typhimurium found deletion in *gipA* gene of L927 and L847 phages. Three complete deletions and four partial deletions in the virulence genes have been found in six of the sequenced strains of *Salmonella enterica*, emphasizing the diverse ecology of phage types. Study on *E. coli* O157:H7 EDL933 strain uncovered various types of indels of IS elements in 12 of their prophages (Iguchi et al., 2006). The prophages of *E. coli* O157 EDL933 have lost their deleterious gene such as λ N gene whose expression would result in host lysis. However, the neutral or beneficial genes for lysogenization like *int* and other phage structural genes occurred in most of their prophages. It was also found that *E. coli* K-12 harbored an incomplete Rac prophage, which was intact in EDL933 strain (Lawrence et al., 2001). A simulation study showed that large-scale deletions occurred in the accessory genes such as integrase and cargo genes, rather than the conserved genes involved in repression of lytic genes (Canchaya et al., 2002), replication, expression of capsid proteins, and packaging. For example, tail protein encoding regions are highly conserved over cargo genes in lambdoid prophages; morons 1 and 2 were least conserved in P2 like prophages (Bobay et al., 2014). The low ion-irradiation experiment showed that the IS elements and pseudogenes as a preferential site for deletions (Song and Luo, 2012) while a study on *Neisseria meningitides* strains showed a IS30 transposase replacement for head and tail morphogenesis genes (Dunning Hotopp et al., 2006). Deletion of inutile parts of the genome confers a growth advantage due to lowered metabolic burden or reduction in the inhibition of growth. Such bacteria will eventually outnumber the counterparts leading to smaller but optimal genomes in the population. However, other views suggest that the degeneration of prophages is driven by the need to eliminate the deleterious genes rather than to preserve chromosome compaction (Lawrence et al., 2001).

WHAT ARE THE ADVANTAGES OF GROUNDED PROPHAGES?

As noted earlier, prophages are dynamic zones within the genome wherein the frequency of genetic variation is very high. There are several disadvantages of bearing the grounded prophages, mostly stemming from the metabolic burden of replicating the large prophage and the expression of prophage genes. However, these disadvantages are subject to contextual economics, i.e., the resources available, growth conditions, and the competition would determine the costs to benefits ratio. The pre-selected lysogens continue to survive unless there are drastic changes in the growth conditions, competition or habitat change. There is also a possibility for the reversal of grounding through homologous recombination of a defective incumbent prophage with a related superinfecting-phage genome, leading to prophage activation and resulting in host lysis. Within a habitat, individual bacteria that may have acquired genetic changes within prophage imposing high costs/benefits are eliminated in the purifying selection. However, this purifying selection based on the prophage would have negligible impact on the species. On the other hand, there is a need to explore various rationales to reason the selection of individuals within a given habitat. There is minimal consistency in the preservation of a prophage. Besides, different prophages are implicated in a plethora of functions to the host genome. Hence, in the following sections we propose various advantages of harboring prophages to explain how prophages may have influenced the bacterial ecological success.

(a) Variations due to Prophage Integration

Every event of integration causes a genetic variation; those that are either tolerable or advantageous are retained, similar to the consequences of transposition. If an integration event were deleterious, the bacterium with such mutations would be eliminated from the population. On the contrary, if a mutation is retained within the population, it must be either tolerable or advantageous. The λ integration at the *exuR* gene, involved in the regulation of *exu* regulon, rendered them non-functional in *E. coli* HfrH 58 strain. It is possible that the expression of *exu* regulon in the ecological context was not required. However, the functionality of *exuR* gene is regained upon prophage induction by site-specific recombination (Mata et al., 1978). Insertional inactivation of lipase-encoding gene by L54a and L54b prophages in *Staphylococcus aureus* PS54 has been reported. However, the curing of L54a prophage has restored the lipase-encoding gene expression but not with L54b curing (Lee and Iandolo, 1985). Similarly, *S. aureus* MW2 harbors ΦSa3mw prophage whose integration happened in haemolytic toxin gene, β-hemolysin (*Hlb*). Upon prophage induction, the *Hlb* gene expression was restored (Katayama et al., 2013). Most of the inversions occur as a result of homologous recombination between prophages (as hot spots for large-scale inversions), IS elements, and RNA-encoding operons (Iguchi et al., 2006; Camilli et al., 2011). Multiple inversions among the Mu-like prophages in various strains are shown to influence the host specificity by altering the expression of various host-cell-receptor recognition genes (Schmucker et al., 1986) and loss of tail genes (Braid et al., 2004). It is indeed difficult to ascertain if any of the phage integrations, solely by virtue of the integration event, has conferred a selective advantage.

(b) Hotspots of Horizontal Gene Transfer

The probability of acquiring exogenous DNA into the genome is dependent on several aspects which include the availability of exogenous DNA, the competence of the host bacterium to take up the DNA, the expression of appropriate recombinases to integrate the DNA, and importantly the costs/benefits of the integration event. There is a lot of evidence that the numerous recombination events in a bacterial genome are associated with their prophages. Multiple prophages encode recombinases, which may aid in integration. However, the prevalence of indels within prophages is not indicative that prophages somehow ‘actively’ acquire genes. Rather, to state accurately, indels may occur randomly in any locus of the bacterial genome, but those recombinant genomes that have resulted in deleterious consequences (high costs) would have been eliminated from the population. Hence, the prevalence of multiple recombination events within the prophages of various strains within a given species implies that these recombinations are either tolerated or advantageous. This aspect could be best explained with the genetic phage mosaicism observed in several prophages. Phages of different groups or broad host range show the shuffling of virulence genes between their genomes (Desiere et al., 2001; Mirold et al., 2001). The number of mosaic phages is higher among gram negative bacteria over gram positive bacteria as the former have a broad host range and are recombinogenic (Brussow and Desiere, 2001).

Lambdoid prophages are capable of supplementing virion proteins to other lambdoid phages (Asadulghani et al., 2009). A study on the genome sequence of HK97, HK022, λ, and P22 prophages has found the mosaic pattern in their functional units such as individual genes, gene clusters, or a portion of protein-encoding genes (Juhala et al., 2000). Recombinases play a vital role in the formation of mosaics among prophages. For example, Rad52-like recombinase promoted the homologous recombination between lambdoid phages and *E. coli* prophage remnants. The prophage remnants in many cases have been found to complement functional genes involved in the lytic cycle. For example, DLP12 lysis protein essential in *E. coli* for the biofilm formation is also found active in lambda phages (De Paepe et al., 2014).

Horizontal gene transfer has led to the integration of ‘morons,’ homologs of bacterial genes, into the phage genomes. The moron gp15 of HK97 prophage in *Enterobacteriaceae* shared no homologs among the other phages but it did with bacterial genes that expressed YebO family of proteins. Other examples include GogB effector protein of *Salmonella enteric* prophage Gifsy-1 (Coombes et al., 2005) and NelA-like type III effector protein of enterohemorrhagic *E. coli* O157:H7 prophage BP4795 (Gruenheid et al., 2004). Similarly, *oac* moron of *Pseudomonas* phage D3 was homologous to *Pseudomonas* WbpC proteins involved in LPS biosynthesis (Burrows et al., 1996).

Prophages may confer direct selective advantages to their host. For example, the lysogenization of *E. coli* by CPS-53 and CP4-57 prophages promote host fitness under oxidative, osmotic, and acidic stresses (Wang et al., 2010). Likewise, RnlA toxin encoding gene in CP4-57 prophage (Koga et al., 2011) and CbtA toxin in CP4-44 (Tan et al., 2011) enhance persister phenotype in *E. coli* by inhibiting the cell growth. The paralogs of *nanS* esterase gene, required for the growth of pathogenic bacteria upon exploiting the host sialic acids, are found downstream of *stx* toxin genes in prophage or prophage remnants of *E. coli* strain O157:H7 (Rangel et al., 2016).

(c) Hotspots of *de novo* Genesis

Since the generation of the genetic information and the evolution of the first cell, most of the genetic information must have evolved within cellular entities. Although it is possible for the *de novo* genesis of genes in bacterial cells, there is minimal tolerance for genetic events within the core gene-rich loci due to high costs and low benefits. The probability and tolerance for recombinations within the prophages serve as a platform for the generation of new genes. Hence, there is a relatively high probability for the *de novo* formation of new open reading frames (Bartels et al., 2009) and possibly selected if the product of the ORF is beneficial for the cell. With time and generations, the *de novo* formed gene may accumulate beneficial mutations and eventually be optimized for an appropriate function and regulated expression. For example, 94.2% of the genome of *Staphylococcus aureus* strain Jevons B contains ϕJB prophage that has 70 predicted ORFs (Bartels et al., 2009) out of which only 21 ORFs were associated with a putative function. The remaining 49 putative proteins have no known function, but structural domains and homologs have been identified for 26 and transmembrane region in 5 ORFs. The remaining proteins were recorded as "hypothetical" which refers to a legitimate ORF but encodes a protein with unknown function or association with any known proteins (Varga et al., 2016). A study on lysogeny module of StB27 prophage among *Staphylococcus hominis* and *Staphylococcus capitis* species found that 42% of the conserved proteins have no predicted functions, and 13 proteins shared no homologs. StB27 of coagulase-negative *Staphylococci* (CoNS) contained ORF3 of unknown function (Deghorain et al., 2012). Several ORFs of

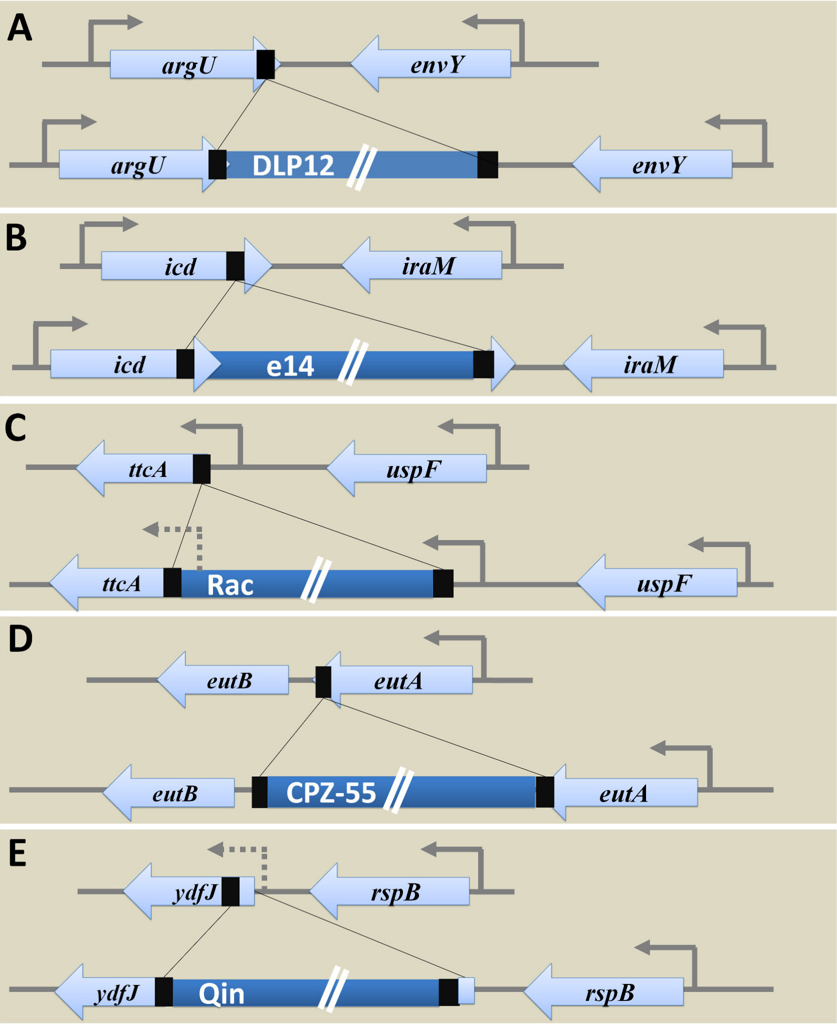
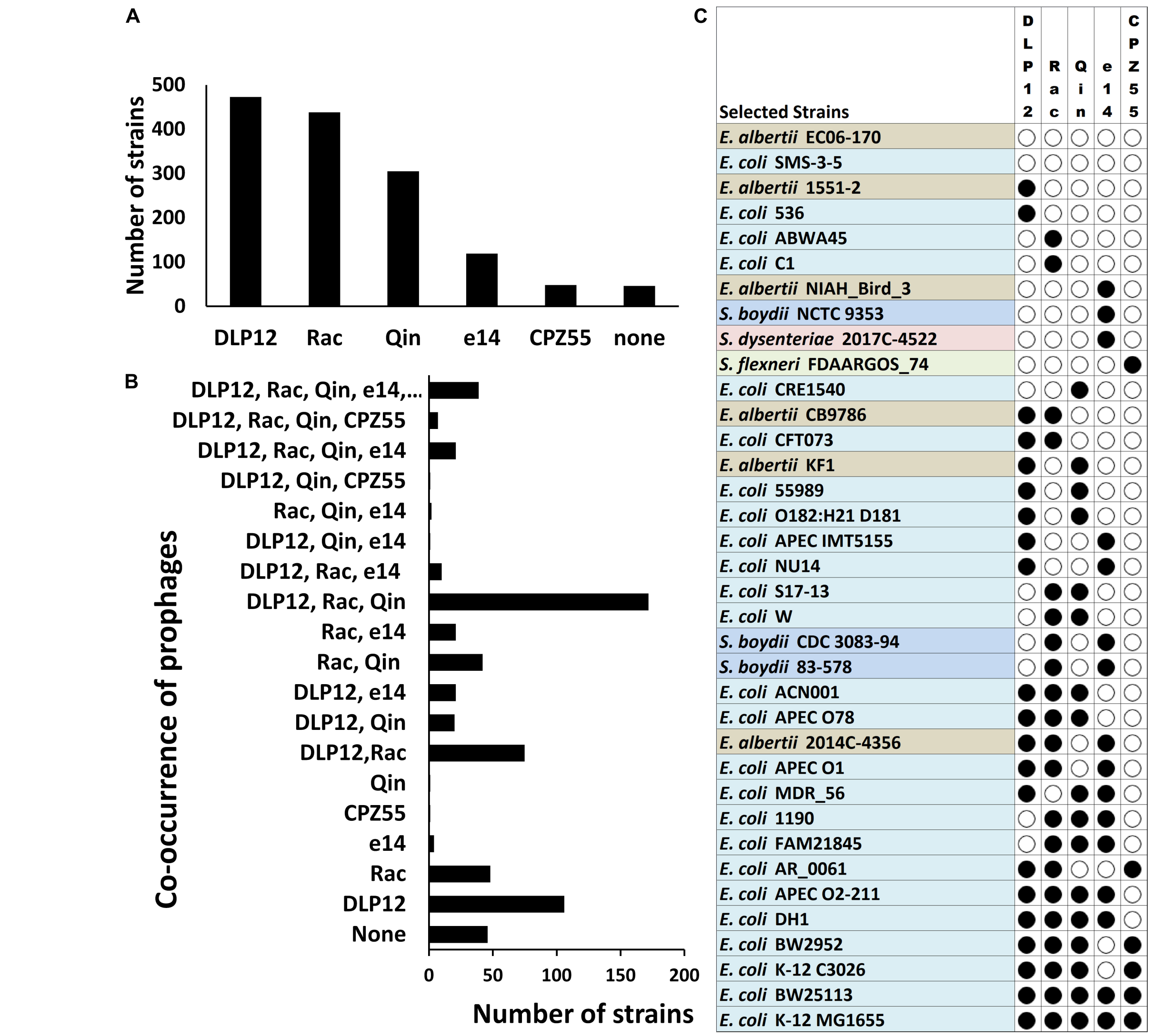


FIGURE 4 | Genetic variations caused by prophage integration in *E. coli* MG1655. The examination of specific integration loci of each of the five prophages (DLP12, e14, Rac, CPZ-55, and Qin) concerning *E. coli* MG1655 among the strains with and without corresponding prophages uncovered the genetic variations caused post integration event as shown above. **(A)** Integration of DLP12 via homologous recombination near the 3′ end of the *argU* gene. The 47 bp region similar to 3′ end of *argU* gene is located on the other end of the DLP12. **(B)** Integration of e14 at the 3′ end of *icd* gene causing the formation of pseudo *icd′* (163 bp) gene as a duplicated region. **(C)** Site-specific integration of Rac at the 5′ end of *ttcA* gene. The variation in the N-terminus of TtcA protein was observed. **(D)** The transposon-like integration of CPZ-55 next to the stop codon of the *eutA* gene causing duplication of 8 bp region. No variations were seen in the bacterial flanking genes (*eutB* and *eutA*). **(E)** The site-specific integration of Qin at the 5′ end of *ydfJ* gene resulting in promoter loss and truncation (lacking 28 codons at the N terminal) of YdfJ protein.



Selected Strains	DLP12	Rac	Qin	e14	CPZ55
E. albertii EC06-170	○	○	○	○	○
E. coli SMS-3-5	○	○	○	○	○
E. albertii 1551-2	●	○	○	○	○
E. coli 536	●	○	○	○	○
E. coli ABWA45	○	●	○	○	○
E. coli C1	○	●	○	○	○
E. albertii NIAH_Bird_3	○	○	○	●	○
S. boydii NCTC 9353	○	○	○	●	○
S. dysenteriae 2017C-4522	○	○	○	●	○
S. flexneri FDAARGOS_74	○	○	○	○	●
E. coli CRE1540	○	○	●	○	○
E. albertii CB9786	●	●	○	○	○
E. coli CFT073	●	●	○	○	○
E. albertii KF1	●	○	●	○	○
E. coli 55989	●	○	●	○	○
E. coli O182:H21 D181	●	○	●	○	○
E. coli APEC IMT5155	●	○	○	●	○
E. coli NU14	●	○	○	●	○
E. coli S17-13	○	●	●	○	○
E. coli W	○	●	●	○	○
S. boydii CDC 3083-94	○	●	○	●	○
S. boydii 83-578	○	●	○	●	○
E. coli ACN001	●	●	●	○	○
E. coli APEC O78	●	●	●	○	○
E. albertii 2014C-4356	●	●	○	●	○
E. coli APEC O1	●	●	○	●	○
E. coli MDR_56	●	○	●	●	○
E. coli 1190	○	●	●	●	○
E. coli FAM21845	○	●	●	●	○
E. coli AR_0061	●	●	○	○	●
E. coli APEC O2-211	●	●	●	●	○
E. coli DH1	●	●	●	●	○
E. coli BW2952	●	●	●	○	●
E. coli K-12 C3026	●	●	●	○	●
E. coli BW25113	●	●	●	●	●
E. coli K-12 MG1655	●	●	●	●	●

FIGURE 5 | Locus-specific distribution of the five prophages (with reference to *E. coli* MG1655 strain) in 638 completely sequenced *E. coli* strains. The cryptic prophages (DLP12, e14, Rac, CPZ-55, and Qin) of MG1655 strain were analyzed for their distribution within 638 completely sequenced *E. coli* strains taking the bacterial and prophage borders sequences as a query in BLASTn. Only cryptic prophages of MG1655 strain with location specificity were analyzed. Other strains may have various other prophages at multiple locations. **(A)** The prevalence of each cryptic prophage within various *Escherichia* and *Shigella* strains. The distribution was determined by curating the BLAST hits obtained in completely sequenced strains showing >85% sequence identity. **(B)** The prevalence of various permutations of prophage co-occurrence within various *Escherichia* and *Shigella* strains. **(C)** The diversity in the locus-specific prophage distribution across *Escherichia* and *Shigella* species. Only representative strains are indicated for illustration purposes.

unknown function and hypothetical proteins have been found in prophages among *E. coli* O157 strains (Svab et al., 2013) (Saile et al., 2016). Similarly, numerous ORFs of unknown functions were identified in multiple species such as *Helicobacter pylori* (Luo et al., 2012), *Ralstonia solanacearum* strains (Askora et al., 2009), and *Pseudomonas aeruginosa* (Braid et al., 2004; Wang et al., 2004). P2-like coliphages of *Enterobacter* have ORFs that have hypothetical protein homologs in plasmids, genomes of other bacterial species (Nilsson et al., 2004). Interestingly, the prophage mEp021 of lysogenic *E. coli* K-12 harbors ORFs whose expression shows haemolytic activity in the host. The ORF-4 has four initiation codons within the single frame encoding for 83aa (ORF-4.1), 82aa (ORF-4.2), 77aa (ORF-4.3), and 72aa (ORF-4.4). Among these, the expression of ORF-4.3 has an inductive role in haemolytic activity by releasing vesicles containing bacterial protein HlyE and alters the morphology of the host bacterium (Martinez-Penafiel et al., 2012). The above examples reiterate that the great extent of the *de novo* genesis of novel genes is relative to any other sites on a typical bacterial genome. However, those ORFs whose products are functional and beneficial will be selected based on the costs/benefits and may evolve into beneficial genes with a high degree of propagative potential.

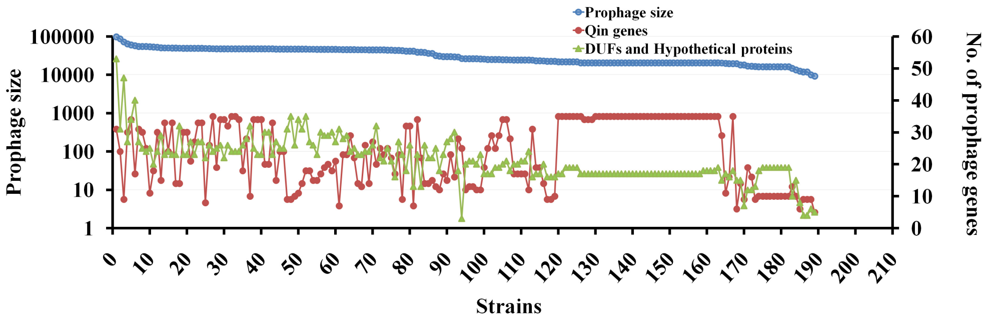


FIGURE 6 | The genetic diversity of Qin prophage in various *E. coli* strains. Correlation between the prophage size and the number of genes carried within the prophage is studied using BLAST tool and curated in excel. For each strain, the length of site- specific Qin prophage and the number of prophage encoding proteins were obtained using BLASTn tool taking gene sequence of each gene as query and hits are curated using excel. The dependence of prophage size and the flux of genes are represented in a scatter plot as shown taking prophage size on primary *Y*-axis, the number of genes on the prophage as secondary *Y*-axis and 189 strains on the *X*-axis. For Qin prophage analyses, the prophage coordinates were used to estimate the prophage size. Having *E. coli* MG1655 as the reference strain, the number of genes carried on Qin prophage among 189 strains was noted. The number of DUFs (Domain of Unknown Function) and hypothetical proteins in Qin prophage of each strain was fetched using a JavaScript to search the GenBank files of each prophage for "DUF" and "hypothetical proteins." The graph was plotted to present the diversity of prophage in terms of its size, the number of genes carried, and novel genes in each of the strains.

GROUNDED PROPHAGES OF *E. coli*: ILLUSTRATIVE ANALYSES

E. coli K12 MG1655 strain has nine cryptic prophages of which DLP12 (20,525 bp) (Lindsey et al., 1989; Toba et al., 2011), Rac (22,890 bp) (Kaiser and Murray, 1979), Qin (20,456 bp) (Kaiser, 1980), e14 (15,204 bp) (Hiom et al., 1991; Mehta et al., 2004), and CPZ-55 (6,802 bp) prophages are well characterized. Several physiological functions such as stress tolerance, biofilm formation, antibiotic resistance, and other advantages to the host genome are attributed to the cryptic prophages in *E. coli* (Wang et al., 2010). These grounded prophages were suggested as potential drug targets (Wang and Wood, 2016). Analysis of locus-specific prophage sequences in completely sequenced *Escherichia* and *Shigella* strains (*E. coli* MG1655 strain as the reference) illustrates the significance of prophages in bacterial genome evolution and ecology at the level of a species. Examining the alignment of the specific locus in strains with and without prophage led to deduce the probable ancestral type locus and decipher the variations caused due to prophage integration. The tRNA genes are hotspots of phage integration based on sequence homology between the phage genome and tRNA gene (Campbell, 2003) and has been shown in case of integration of DLP12 prophage near the 3′ end of *argU* gene encoding tRNA-Arg. The probable recombinant *attR* site is similar to 47 bp of the *argU* gene, which was located on the other end of the prophage. The site-specific integration of e14 within the *icd* gene coding for isocitrate dehydrogenase protein (416aa) separated the 163 bp region from 3′ end of *icd* transcript. The integrated prophage provides the 3' terminal codon of the *icd* ORF and thus forming homologous *attL* and *attR* sites [**Supplementary Table S1** (**Datasheet S1**)]. Hence, the excision of e14 prophage results in restoration of functional *icd* gene yet rendering conservative replacement of aspartate by glutamate encoding codon (Hill et al., 1989). The site-specific integration of Rac prophage at the 5′ end of *ttcA* gene via the *attB* and *attP* sites resulted in the recombinant *attL* and *attR* sites. Variations in the amino acid sequence of the N-terminus of TtcA protein [a 311aa tRNA-cytidine(32) 2-sulfurtransferase protein involved in post-transcriptional thiolation of tRNA] was observed within the strains harboring Rac prophage. Such variations owing to sequence dissimilarity between the recombinant *attR* and *attL* sites reduce the fitness level in carbenicillin positive niche. The excision of Rac enhances biofilm formation, cell lysis, and motility (Hong et al., 2010; Liu et al., 2015). Integration of CPZ-55 prophage has occurred right after the stop codon of the *eutA* gene coding for ethanolamine utilization protein, causing target site duplication of 8 bp sequence (5′-TCAGGAAG-3′). The intergenic region of 13 bp (5′-TAAGTCGTTCCCT-3′) was conserved. No variation observed in prophage flanking genes, *eutB* (453 codons) and *eutA* (467 codons). However, CPZ-55 integration has separated the genes *eutA* and *eutB* and thereby disrupting the functionality of *eut* operon encoding proteins such as EutS, EutQ, EutN, EutD, EutM, EutE, EutH, and others, that are involved in ethanolamine reduction followed by utilization when available as the only nitrogen source. As the initial steps in the utilization of ethanolamine requires the expression of *eutA*, *eutB*, and *eutC*, any alterations as caused by prophage integration would result in the inability of *eut*$^{-}$ strains to multiply upon utilizing ethanolamine (Soupene et al., 2003) (**Figure 4**) [**Supplementary Table S1** (**Datasheet S1**)].

We analyzed the distribution of the selected cryptic prophages within different *Escherichia* and *Shigella* strains varies. Locus-specific sequence analysis of the above mentioned five prophages (loci as per *E. coli* MG1655) within the completely sequenced *Escherichia* and *Shigella* strains (a total of 638 strains) shows that DLP12 is the most prevalent (74%) prophage followed by Rac prophage in 69% of total strains (**Figure 5A**). Qin prophage is present in 48%, and e14 prophage is encoded by 19%, while only 8% of total strains encoded CPZ-55 prophage. The prevalence in the *E. coli* genomes indicates that the integration of DLP12 and Rac have happened earlier relative to the other prophages analyzed here. Forty-six strains did not harbor any of the five prophages scored for. The analysis of co-occurrence of prophages within *Escherichia* and *Shigella* strains would shed some light on the acquisition of the prophages during the evolution of these bacteria. Hence, we analyzed the co-occurrence of the five prophages (DLP12, e14, Rac, CPZ-55 and Qin) with all possible permutations in 638 completely sequenced and annotated *Escherichia* and *Shigella* genomes (available as on March 2018) (**Figure 5**). Although there are 32 different permutations of the five prophages, only 19 combinations were observed in our analysis [**Supplementary Table S3** (**Datasheet S2**)]. If prophages were acquired sequentially, it would be expected that the various observed permutations would have an incremental pattern of prophages within various genomes. 172 *E. coli* genomes had DLP12, Rac, and Qin cryptic prophages while 106 genomes had DLP12 prophage and none of the other four prophages. Thirty-nine *E. coli* genomes including most of the K12 strains had the DLP12, Rac, Qin, e14, and CPZ-55 permutation (**Figure 5B**). In essence, we did not notice any particular pattern that might imply serial acquisition of cryptic prophages. For example, DLP12 is present with various combinations of other prophages that were analyzed indicating random co-occurrence (**Figure 5C**). The observed permutations are mostly random which would mean that these cryptic prophages also propagated horizontally among the *E. coli* strains. Horizontal propagation is common among bacterial strains; however, the mechanism of HGT is most likely through conjugation considering the large size of the cryptic prophages.

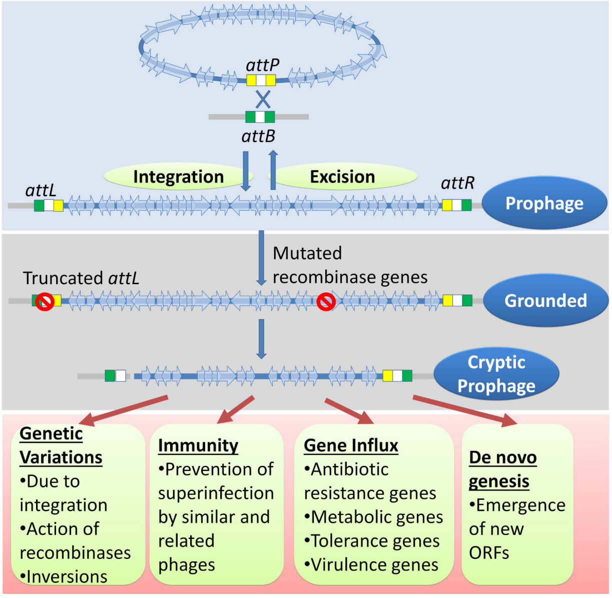


FIGURE 7 | A genetic history and advantages of cryptic prophages. The genome of temperate phages integrates with the bacterial chromosome through site-specific recombination between sites similar to those of *attP* and *attB*. Occasionally, mutations in the *attL*, *attR*, recombinase genes, and genes encoding other essential accessory factors would result in a 'grounded' prophage. Eventually but gradually, more mutations, usually deletions, accumulate to ease the metabolic and genetic burden imposed by the cryptic prophage. The cryptic prophage offers several advantages to the host bacterium. The genetic variations caused in the host genome due to integrations, inversions involving prophage sequences, and the action of their recombinases may have a selective advantage. The prophage confers immunity from secondary phage infection and acts as hotspots for gene influx through horizontal gene transfer of important genes such as those involved in metabolism, antibiotic resistance, virulence, and stress tolerance. More importantly, the cryptic prophage is a site for the emergence of new ORFs that may have novel and useful functions.

From our analysis of the 36 genes of Qin prophage from *E. coli* MG1655 strain, genes like *hokD*, *relBE*, cold-shock genes (e.g., *cspI*, *cspB*, etc.), and regulators (e.g., *dicC*, *dicB*, etc.) were predicted to have been acquired via HGT [**Supplementary Table S2** (**Datasheet S1**)]. In our sequence analyses, we found multiple mutations in *intQ'* of most Qin prophage hits, which is indicative that the inactivation of the integrase gene could be a reason for the grounding of the Qin prophage. Although several Qin prophage genes are traditionally associated with general phage genes, numerous genes associated with Qin prophages are obtained by HGT. Several genes from plasmids and bacterial genomes (predominantly from *E. coli*, *E. albertii*, and *Citrobacter*) have integrated into the Qin prophage. The ancestral *hokD* ORF which is about 70 codons, however, is only 51 codons due to the integration of the *relBE* Toxin-Antitoxin system which is acquired from plasmids (Ramisetty and Santhosh, 2016). This observation emphasizes cautions to be exercised during characterization of genes that are highly linked to or within prophages because of high probability for genetic alterations. Comparisons of sequences of the same gene across several strains should be made to determine any truncations or modifications induced due to multiple genetic changes.

We then performed a comparative analysis of Qin prophage diversity within various *Escherichia* strains by estimating the length of prophage in base pairs (distance between the *att* sites), the number of Qin genes (with reference to MG1655 strain), and the number of *de novo* ORFs using a semantic script (hypothetical genes and DUFs). The size of the Qin prophage in 189 strains was highly divergent ranging from 11,280 bp in *E. coli* O6:H6 strain M9682-C1 to 97,450 bp in *E. coli* O26:H11 str. 11368 DNA (**Figure 6**). Surprisingly, we noticed a slight variation in Qin prophage size (of difference ∼1,300 bp) even within K-12 strains, which is indicative that prophages are highly prone to mutations. The number of copies of phage genes encoding major and minor capsid proteins, tail measure protein, host specificity protein, transposase, and other phage-related genes increased with the size of the Qin prophage. This observation indicates that cryptic prophages are hotspots for superintegration of phage/plasmid genomes (Lichens-Park et al., 1990). Conceivably, a positive correlation between the size of the prophage and the number of hypothetical proteins can be found. Some of the hypothetical genes could have accumulated as more plasmid/phage genes integrated into the Qin prophage. Nevertheless, there is a high probability for the generation of ORFs *de novo* due to multiple recombination events (Delaye et al., 2008; Tautz and Domazet-Lošo, 2011). Although the expressivity and functionality of most of the genes that originated through *de novo* genesis are unclear, they are genes in the making which maybe already functional or are yet to evolve for a beneficial function in the future generations.

To predict the possible role of the cryptic prophage in contributing to the ecological advantages to the host bacterium, we manually scouted for genes that are unrelated to the phage/plasmid in the Qin prophages of various *E. coli* strains [**Supplementary Table S3** (**Datasheet S2**)]. The Qin prophage from strains examined carried diverse genes involved in significant functions such as transportation (e.g., Calcium transporter (*chaC*) in *E. coli* strain CRE1540), metabolism (e.g., Acyl-CoA thioesterase in *E. coli* O111:H-str. 11128 DNA), recombination (e.g., RusA endodeoxyribonuclease in *E. coli* strain ETEC-2265), virulence (e.g., Shiga toxin Stx2 subunits A and B in *E. coli* O178:H19 strain 2012C-4431), transcriptional regulation (e.g., LysR transcriptional regulator in *E. coli* strain CRE1540), and resistance (e.g., Tellurite/Colicin resistance protein in *E. coli* SE11 DNA) [**Supplementary Table S3** (**Datasheet S2**)]. Using a semantic script, a total of 4,034 hypothetical proteins/DUFs were found in Qin prophages of 189 *Escherichia* and *Shigella* genomes indicating an average of 21 hypothetical genes per Qin prophage. Even after discounting 50% of the hypothetical genes for annotation and functionality issues, the high density and diversity of the hypothetical genes per prophage is indicative that grounded prophages are sites of *de novo* genesis which could expedite genome diversification and evolution (Berglund et al., 2010). In essence, grounded prophages of *E. coli* are hotspots of multiple recombination events, genetic flux, and *de novo* genesis of genes which expedite adaptations to various habitats.

CONCLUSION

The physiology of higher organisms is vastly influenced by their respective microbiomes. Microbiomes, in turn, are influenced by their phages (Scanlan, 2017; De Sordi et al., 2018). Bacteriophages are the major biological drivers of bacterial ecology and evolution through strategies such as symbiosis, dependency, and dormancy (Roossinck, 2011; Nasir et al., 2017). The prevalence of multiple grounded prophages within most bacterial genomes implies that the sequence of events such as lysogenization, grounding of prophages, domestication of phage genes, and the genetic events within the grounded prophages have an evolutionary impact on the selection of bacterial genomes. (1) Lysogeny, as opposed to the lytic cycle, represents an 'evolutionarily stable strategy' between bacteria and phages growing in an econiche to allow the existence of both the phage and the host. (2) The selection of lysogens is favored by the elimination of intraspecies competition by the phage and the immunity conferred to lysogens by the prophages. (3) The lysogens with grounded prophages are selected over lysogens (with wild-type prophages) because the latter have an imminent threat of prophage activation followed by lysis. The grounded prophages may provide immunity from specific phage attack, ecologically significant traits, and the evolution of new genes (**Figure 7**). Many genes that provide stress tolerance, antibiotic resistance, virulence genes, and metabolic genes acquired horizontally are located within prophages. (4) Grounded prophages are zones within genomes that allow multiple recombinations and mutations with minimal deleterious effects on the host physiology. The increased probability of genetic modifications (insertions, deletions, and inversions) within the prophages increases the probability of formation of new ORFs. These *de novo* genes maybe refined gradually for a function with a selective advantage such as antibiotic resistance and virulence.

Since the evolution of the first temperate phages, prophages must have enhanced the rate of evolution in different life forms including higher organisms. Grounded prophages are the most significant drivers of ecology, genetic diversification, and microbial genome evolution. The grounded prophages, and the analogous elements in other cellular life-forms, are the 'holy grails' of *de novo* genesis of genetic information, which must be of immense interest to molecular evolutionists.

AUTHOR CONTRIBUTIONS

BCMR has conceived the idea, designed the work, performed the analyses, interpreted the results, and wrote the manuscript. PAS has performed the work, interpreted the results, and wrote the manuscript.

FUNDING

The work in BCMR's lab is supported by Prof. T. R. Rangarajan grants (internal funding) and the Central Research Facility provided by SASTRA Deemed University, Thanjavur, India.

ACKNOWLEDGMENTS

The authors thank Sri Uma Aisvarya Balasubramanian for proofreading the manuscript. The authors also thank the reviewers for their valuable suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00065/full#supplementary-material

REFERENCES

Abe, K., Kawano, Y., Iwamoto, K., Arai, K., Maruyama, Y., Eichenberger, P., et al. (2014). Developmentally-regulated excision of the SPbeta prophage reconstitutes a gene required for spore envelope maturation in *Bacillus subtilis*. *PLoS Genet.* 10:e1004636. doi: 10.1371/journal.pgen.1004636

Allet, B. (1979). Mu insertion duplicates a 5 base pair sequence at the host inserted site. *Cell* 16, 123–129. doi: 10.1016/0092-8674(79)90193-4

Asadulghani, M., Ogura, Y., Ooka, T., Itoh, T., Sawaguchi, A., Iguchi, A., et al. (2009). The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* 5:e1000408. doi: 10.1371/journal.ppat.1000408

Askora, A., Kawasaki, T., Usami, S., Fujie, M., and Yamada, T. (2009). Host recognition and integration of filamentous phage phiRSM in the phytopathogen, *Ralstonia solanacearum*. *Virology* 384, 69–76. doi: 10.1016/j.virol.2008.11.007

Bartels, M. D., Boye, K., Rohde, S. M., Larsen, A. R., Torfs, H., Bouchy, P., et al. (2009). A common variant of staphylococcal cassette chromosome mec type IVa in isolates from Copenhagen, Denmark, is not detected by the BD GeneOhm methicillin-resistant *Staphylococcus aureus* assay. *J. Clin. Microbiol.* 47, 1524–1527. doi: 10.1128/JCM.02153-08

Berglund, E. C., Ellegaard, K., Granberg, F., Xie, Z., Maruyama, S., Kosoy, M. Y., et al. (2010). Rapid diversification by recombination in *Bartonella grahamii* from wild rodents in Asia contrasts with low levels of genomic divergence in Northern Europe and America. *Mol. Ecol.* 19, 2241–2255. doi: 10.1111/j.1365-294X.2010.04646.x

Betts, A., Gray, C., Zelek, M., Maclean, R. C., and King, K. C. (2018). High parasite diversity accelerates host adaptation and diversification. *Science* 360, 907–911. doi: 10.1126/science.aam9974

Bibb, L. A., Hancox, M. I., and Hatfull, G. F. (2005). Integration and excision by the large serine recombinase phiRv1 integrase. *Mol. Microbiol.* 55, 1896–1910. doi: 10.1111/j.1365-2958.2005.04517.x

Blackstone, N. W., and Green, D. R. (1999). The evolution of a mechanism of cell suicide. *Bioessays* 21, 84–88. doi: 10.1002/(SICI)1521-1878(199901)21:1<84::AID-BIES11>3.0.CO;2-0

Blokesch, M. (2017). In and out-contribution of natural transformation to the shuffling of large genomic regions. *Curr. Opin. Microbiol.* 38, 22–29. doi: 10.1016/j.mib.2017.04.001

Bobay, L. M., Rocha, E. P., and Touchon, M. (2013). The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* 30, 737–751. doi: 10.1093/molbev/mss279

Bobay, L. M., Touchon, M., and Rocha, E. P. (2014). Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12127–12132. doi: 10.1073/pnas.1405336111

Bonneau, R., Baliga, N. S., Deutsch, E. W., Shannon, P., and Hood, L. (2004). Comprehensive de novo structure prediction in a systems-biology context for the archaea *Halobacterium* sp. NRC-1. *Genome Biol.* 5:R52. doi: 10.1186/gb-2004-5-8-r52

Braga, L. P. P., Soucy, S. M., Amgarten, D. E., Da Silva, A. M., and Setubal, J. C. (2018). Bacterial diversification in the light of the interactions with phages: the genetic symbionts and their role in ecological speciation. *Front. Ecol. Evol.* 6:6. doi: 10.3389/fevo.2018.00006

Braid, M. D., Silhavy, J. L., Kitts, C. L., Cano, R. J., and Howe, M. M. (2004). Complete genomic sequence of bacteriophage B3, a Mu-like phage of *Pseudomonas aeruginosa*. *J. Bacteriol.* 186, 6560–6574. doi: 10.1128/JB.186.19.6560-6574.2004

Brody, H., Greener, A., and Hill, C. W. (1985). Excision and reintegration of the *Escherichia coli* K-12 chromosomal element e14. *J. Bacteriol.* 161, 1112–1117.

Brody, H., and Hill, C. W. (1988). Attachment site of the genetic element e14. *J. Bacteriol.* 170, 2040–2044. doi: 10.1128/jb.170.5.2040-2044.1988

Brussow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602. doi: 10.1128/MMBR.68.3.560-602.2004

Brussow, H., and Desiere, F. (2001). Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Mol. Microbiol.* 39, 213–222. doi: 10.1046/j.1365-2958.2001.02228.x

Buckling, A., and Brockhurst, M. (2012). Bacteria-virus coevolution. *Adv. Exp. Med. Biol.* 751, 347–370. doi: 10.1007/978-1-4614-3567-9_16

Burrows, L. L., Charter, D. F., and Lam, J. S. (1996). Molecular characterization of the *Pseudomonas aeruginosa* serotype O5 (PAO1) B-band lipopolysaccharide gene cluster. *Mol. Microbiol.* 22, 481–495. doi: 10.1046/j.1365-2958.1996.1351503.x

Camilli, R., Bonnal, R. J., Del Grosso, M., Iacono, M., Corti, G., Rizzi, E., et al. (2011). Complete genome sequence of a serotype 11A, ST62 *Streptococcus pneumoniae* invasive isolate. *BMC Microbiol.* 11:25. doi: 10.1186/1471-2180-11-25

Campbell, A. (2003). Prophage insertion sites. *Res. Microbiol.* 154, 277–282. doi: 10.1016/S0923-2508(03)00071-8

Canchaya, C., Desiere, F., Mcshan, W. M., Ferretti, J. J., Parkhill, J., and Brussow, H. (2002). Genome analysis of an inducible prophage and prophage remnants integrated in the *Streptococcus pyogenes* strain SF370. *Virology* 302, 245–258. doi: 10.1006/viro.2002.1570

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol. Mol. Biol. Rev.* 67, 238–276. doi: 10.1128/MMBR.67.2.238-276.2003

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* 49, 277–300. doi: 10.1046/j.1365-2958.2003.03580.x

Chakraborty, D., Clark, E., Mauro, S. A., and Koudelka, G. B. (2018). Molecular mechanisms governing "hair-trigger" induction of shiga toxin-encoding prophages. *Viruses* 10:E228. doi: 10.3390/v10050228

Coombes, B. K., Wickham, M. E., Brown, N. F., Lemire, S., Bossi, L., Hsiao, W. W., et al. (2005). Genetic and molecular analysis of GogB, a phage-encoded type III-secreted substrate in *Salmonella enterica* serovar typhimurium with autonomous expression from its associated phage. *J. Mol. Biol.* 348, 817–830. doi: 10.1016/j.jmb.2005.03.024

Craig, N. L., and Nash, H. A. (1983). The mechanism of phage lambda site-specific recombination: site-specific breakage of DNA by Int topoisomerase. *Cell* 35, 795–803. doi: 10.1016/0092-8674(83)90112-5

Cumby, N., Edwards, A. M., Davidson, A. R., and Maxwell, K. L. (2012). The bacteriophage HK97 gp15 moron element encodes a novel superinfection exclusion protein. *J. Bacteriol.* 194, 5012–5019. doi: 10.1128/JB.00843-12

Darmon, E., and Leach, D. R. (2014). Bacterial genome instability. *Microbiol. Mol. Biol. Rev.* 78, 1–39. doi: 10.1128/MMBR.00035-13

Davison, J. (1999). Genetic exchange between bacteria in the environment. *Plasmid* 42, 73–91. doi: 10.1006/plas.1999.1421

Delaye, L., Deluna, A., Lazcano, A., and Becerra, A. (2008). The origin of a novel gene through overprinting in *Escherichia coli*. *BMC Evol. Biol.* 8:31. doi: 10.1186/1471-2148-8-31

De Paepe, M., Hutinet, G., Son, O., Amarir-Bouhram, J., Schbath, S., and Petit, M. A. (2014). Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. *PLoS Genet.* 10:e1004181. doi: 10.1371/journal.pgen.1004181

De Sordi, L., Lourenco, M., and Debarbieux, L. (2018). "I will survive": a tale of bacteriophage-bacteria coevolution in the gut. *Gut Microbes* 10, 92–99. doi: 10.1080/19490976.2018.1474322

Deghorain, M., Bobay, L. M., Smeesters, P. R., Bousbata, S., Vermeersch, M., Perez-Morga, D., et al. (2012). Characterization of novel phages isolated in coagulase-negative staphylococci reveals evolutionary relationships with *Staphylococcus aureus* phages. *J. Bacteriol.* 194, 5829–5839. doi: 10.1128/JB.01085-12

Desiere, F., Mcshan, W. M., Van Sinderen, D., Ferretti, J. J., and Brussow, H. (2001). Comparative genomics reveals close genetic relationships between phages from dairy bacteria and pathogenic *Streptococci*: evolutionary implications for prophage-host interactions. *Virology* 288, 325–341. doi: 10.1006/viro.2001.1085

Dobrindt, U., and Hacker, J. (2001). Whole genome plasticity in pathogenic bacteria. *Curr. Opin. Microbiol.* 4, 550–557. doi: 10.1016/S1369-5274(00)00250-2

Dobrindt, U., Zdziarski, J., Salvador, E., and Hacker, J. (2010). Bacterial genome plasticity and its impact on adaptation during persistent infection. *Int. J. Med. Microbiol.* 300, 363–366. doi: 10.1016/j.ijmm.2010.04.010

Dorgai, L., Oberto, J., and Weisberg, R. A. (1993). Xis and Fis proteins prevent site-specific DNA inversion in lysogens of phage HK022. *J. Bacteriol.* 175, 693–700. doi: 10.1128/jb.175.3.693-700.1993

D'Souza, G., Waschina, S., Kaleta, C., and Kost, C. (2015). Plasticity and epistasis strongly affect bacterial fitness after losing multiple metabolic genes. *Evolution* 69, 1244–1254. doi: 10.1111/evo.12640

Dunning Hotopp, J. C., Grifantini, R., Kumar, N., Tzeng, Y. L., Fouts, D., Frigimelica, E., et al. (2006). Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes. *Microbiology* 152, 3733–3749. doi: 10.1099/mic.0.29261-0

Durand, P. M., Sym, S., and Michod, R. E. (2016). Programmed cell death and complexity in microbial systems. *Curr. Biol.* 26, R587–R593. doi: 10.1016/j.cub.2016.05.057

Dydecka, A., Bloch, S., Rizvi, A., Perez, S., Nejman-Falenczyk, B., Topka, G., et al. (2017). Bad phages in good bacteria: role of the mysterious orf63 of lambda and shiga toxin-converting Phi24B Bacteriophages. *Front. Microbiol.* 8:1618. doi: 10.3389/fmicb.2017.01618

Farrant, J. L., Sansone, A., Canvin, J. R., Pallen, M. J., Langford, P. R., Wallis, T. S., et al. (1997). Bacterial copper- and zinc-cofactored superoxide dismutase contributes to the pathogenesis of systemic salmonellosis. *Mol. Microbiol.* 25, 785–796. doi: 10.1046/j.1365-2958.1997.5151877.x

Fierer, J., and Guiney, D. G. (2001). Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection. *J. Clin. Invest.* 107, 775–780. doi: 10.1172/JCI12561

Finkel, S. E., and Kolter, R. (2001). DNA as a nutrient: novel role for bacterial competence gene homologs. *J. Bacteriol.* 183, 6288–6293. doi: 10.1128/JB.183.21.6288-6293.2001

Fischer, W., Hofreuter, D., and Haas, R. (2001). "Natural transformation, recombination, and repair," in *Helicobacter pylori: Physiology and Genetics,* eds H. L. T. Mobley, G. L. Mendz, and S. L. Hazell (Washington, DC: ASM Press).

Fogg, P. C. M., Allison, H. E., Saunders, J. R., and Mccarthy, A. J. (2010). *Bacteriophage lambda*: a paradigm revisited. *J. Virol.* 84, 6876–6879. doi: 10.1128/JVI.02177-09

Forde, S. E., Thompson, J. N., and Bohannan, B. J. (2004). Adaptation varies through space and time in a coevolving host-parasitoid interaction. *Nature* 431, 841–844. doi: 10.1038/nature02906

Fujiwara, A., Kawasaki, T., Usami, S., Fujie, M., and Yamada, T. (2008). Genomic characterization of *Ralstonia solanacearum* phage phiRSA1 and its related prophage (phiRSX) in strain GMI1000. *J. Bacteriol.* 190, 143–156. doi: 10.1128/JB.01158-07

Gandon, S. (2016). Why be temperate: lessons from *Bacteriophage lambda*. *Trends Microbiol.* 24, 356–365. doi: 10.1016/j.tim.2016.02.008

Gaudriault, S., Duchaud, E., Lanois, A., Canoy, A. S., Bourot, S., Derose, R., et al. (2006). Whole-genome comparison between *Photorhabdus* strains to identify genomic regions involved in the specificity of nematode interaction. *J. Bacteriol.* 188, 809–814. doi: 10.1128/JB.188.2.809-814.2006

Ghosh, P., Wasil, L. R., and Hatfull, G. F. (2006). Control of phage Bxb1 excision by a novel recombination directionality factor. *PLoS Biol.* 4:e186. doi: 10.1371/journal.pbio.0040186

Gottesman, M. E., and Weisberg, R. A. (2004). Little lambda, who made thee? *Microbiol. Mol. Biol. Rev.* 68, 796–813. doi: 10.1128/MMBR.68.4.796-813.2004

Groth, A. C., and Calos, M. P. (2004). Phage integrases: biology and applications. *J. Mol. Biol.* 335, 667–678. doi: 10.1016/j.jmb.2003.09.082

Gruenheid, S., Sekirov, I., Thomas, N. A., Deng, W., O'donnell, P., Goode, D., et al. (2004). Identification and characterization of NleA, a non-LEE-encoded type III translocated virulence factor of enterohaemorrhagic *Escherichia coli* O157:H7. *Mol. Microbiol.* 51, 1233–1249. doi: 10.1046/j.1365-2958.2003.03911.x

Hacker, J., and Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* 2, 376–381. doi: 10.1093/embo-reports/kve097

Hargreaves, K. R., Kropinski, A. M., and Clokie, M. R. (2014). What does the talking?: Quorum sensing signalling genes discovered in a bacteriophage genome. *PLoS One* 9:e85131. doi: 10.1371/journal.pone.0085131

Harshey, R. M. (2014). Transposable phage Mu. *Microbiol. Spectr.* 2, 669–691. doi: 10.1128/microbiolspec.MDNA3-0007-2014

Hill, C. W., Gray, J. A., and Brody, H. (1989). Use of the isocitrate dehydrogenase structural gene for attachment of e14 in *Escherichia coli* K-12. *J. Bacteriol.* 171, 4083–4084. doi: 10.1128/jb.171.7.4083-4084.1989

Hiom, K., Thomas, S. M., and Sedgwick, S. G. (1991). Different mechanisms for SOS induced alleviation of DNA restriction in *Escherichia coli*. *Biochimie* 73, 399–405. doi: 10.1016/0300-9084(91)90106-B

Hoess, R. H., and Abremski, K. (1985). Mechanism of strand cleavage and exchange in the Cre-lox site-specific recombination system. *J. Mol. Biol.* 181, 351–362. doi: 10.1016/0022-2836(85)90224-4

Hogardt, M., Hoboth, C., Schmoldt, S., Henke, C., Bader, L., and Heesemann, J. (2007). Stage-specific adaptation of hypermutable *Pseudomonas aeruginosa* isolates during chronic pulmonary infection in patients with cystic fibrosis. *J. Infect. Dis.* 195, 70–80. doi: 10.1086/509821

Hong, S. H., Wang, X., and Wood, T. K. (2010). Controlling biofilm formation, prophage excision and cell death by rewiring global regulator H-NS of *Escherichia coli*. *Microb. Biotechnol.* 3, 344–356. doi: 10.1111/j.1751-7915.2010.00164.x

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., and Sullivan, M. B. (2017). Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J.* 11, 1511–1520. doi: 10.1038/ismej.2017.16

Hurwitz, B. L., Brum, J. R., and Sullivan, M. B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J.* 9, 472–484. doi: 10.1038/ismej.2014.143

Hyder, S. L., and Streitfeld, M. M. (1978). Transfer of erythromycin resistance from clinically isolated lysogenic strains of *Streptococcus pyogenes* via their endogenous phage. *J. Infect. Dis.* 138, 281–286. doi: 10.1093/infdis/138.3.281

Iguchi, A., Iyoda, S., Terajima, J., Watanabe, H., and Osawa, R. (2006). Spontaneous recombination between homologous prophage regions causes large-scale inversions within the *Escherichia coli* O157:H7 chromosome. *Gene* 372, 199–207. doi: 10.1016/j.gene.2006.01.005

Iversen, H., L'Abée-Lund, T. M., Aspholm, M., Arnesen, L. P., and Lindback, T. (2015). Commensal *E. coli* Stx2 lysogens produce high levels of phages after spontaneous prophage induction. *Front. Cell. Infect. Microbiol.* 5:5. doi: 10.3389/fcimb.2015.00005

Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F., and Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J. Mol. Biol.* 299, 27–51. doi: 10.1006/jmbi.2000.3729

Kaiser, K. (1980). The origin of Q-independent derivatives of phage lambda. *Mol. Gen Genet.* 179, 547–554. doi: 10.1007/BF00271744

Kaiser, K., and Murray, N. E. (1979). Physical characterisation of the "Rac prophage" in *E. coli* K12. *Mol. Gen. Genet.* 175, 159–174. doi: 10.1007/BF00425532

Katayama, Y., Baba, T., Sekine, M., Fukuda, M., and Hiramatsu, K. (2013). Beta-hemolysin promotes skin colonization by *Staphylococcus aureus*. *J. Bacteriol.* 195, 1194–1203. doi: 10.1128/JB.01786-12

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

Koga, M., Otsuka, Y., Lemire, S., and Yonesaki, T. (2011). *Escherichia coli* rnlA and rnlB compose a novel toxin-antitoxin system. *Genetics* 187, 123–130. doi: 10.1534/genetics.110.121798

Koskella, B., and Brockhurst, M. A. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* 38, 916–931. doi: 10.1111/1574-6976.12072

Kraft, C., Stack, A., Josenhans, C., Niehus, E., Dietrich, G., Correa, P., et al. (2006). Genomic changes during chronic *Helicobacter pylori* infection. *J. Bacteriol.* 188, 249–254. doi: 10.1128/JB.188.1.249-254.2006

Kutsukake, K., and Iino, T. (1980). A trans-acting factor mediates inversion of a specific DNA segment in flagellar phase variation of *Salmonella*. *Nature* 284, 479–481. doi: 10.1038/284479a0

Landy, A., and Ross, W. (1977). Viral integration and excision: structure of the lambda att sites: DNA sequences have been determined for regions involved in lambda site-specific recombination. *Science (New York, N.Y.)* 197, 1147–1160. doi: 10.1126/science.331474

Lawrence, J. G., Hendrix, R. W., and Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? *Trends Microbiol.* 9, 535–540. doi: 10.1016/S0966-842X(01)02198-9

Lee, C. Y., and Iandolo, J. J. (1985). Mechanism of bacteriophage conversion of lipase activity in *Staphylococcus aureus*. *J. Bacteriol.* 164, 288–293.

Lichens-Park, A., Smith, C. L., and Syvanen, M. (1990). Integration of bacteriophage lambda into the cryptic lambdoid prophages of *Escherichia coli*. *J. Bacteriol.* 172, 2201–2208. doi: 10.1128/jb.172.5.2201-2208.1990

Lindsey, D. F., Mullin, D. A., and Walker, J. R. (1989). Characterization of the cryptic lambdoid prophage DLP12 of *Escherichia coli* and overlap of the DLP12 integrase gene with the tRNA gene argU. *J. Bacteriol.* 171, 6197–6205. doi: 10.1128/jb.171.11.6197-6205.1989

Liu, X., Li, Y., Guo, Y., Zeng, Z., Li, B., Wood, T. K., et al. (2015). Physiological function of rac prophage during biofilm formation and regulation of rac excision in *Escherichia coli* K-12. *Sci. Rep.* 5:16074. doi: 10.1038/srep16074

Lopez, C. A., Winter, S. E., Rivera-Chavez, F., Xavier, M. N., Poon, V., Nuccio, S. P., et al. (2012). Phage-mediated acquisition of a type III secreted effector protein boosts growth of *Salmonella* by nitrate respiration. *MBio* 3:e00143-12. doi: 10.1128/mBio.00143-12

Los, J., Los, M., Wegrzyn, A., and Wegrzyn, G. (2013). Altruism of Shiga toxin-producing *Escherichia coli*: recent hypothesis versus experimental results. *Front. Cell. Infect. Microbiol.* 2:166. doi: 10.3389/fcimb.2012.00166

Los, J. M., Los, M., Wegrzyn, A., and Wegrzyn, G. (2010). Hydrogen peroxide-mediated induction of the Shiga toxin-converting lambdoid prophage ST2-8624 in *Escherichia coli* O157:H7. *FEMS Immunol. Med. Microbiol.* 58, 322–329. doi: 10.1111/j.1574-695X.2009.00644.x

Luo, C. H., Chiou, P. Y., Yang, C. Y., and Lin, N. T. (2012). Genome, integration, and transduction of a novel temperate phage of *Helicobacter pylori*. *J. Virol.* 86, 8781–8792. doi: 10.1128/JVI.00446-12

Mahony, J., Mcgrath, S., Fitzgerald, G. F., and Van Sinderen, D. (2008). Identification and characterization of lactococcal-prophage-carried superinfection exclusion genes. *Appl. Environ. Microbiol.* 74, 6206–6215. doi: 10.1128/AEM.01053-08

Martinez-Penafiel, E., Fernandez-Ramirez, F., Ishida, C., Reyes-Cortes, R., Sepulveda-Robles, O., Guarneros-Pena, G., et al. (2012). Overexpression of Ipe protein from the coliphage mEp021 induces pleiotropic effects involving haemolysis by HlyE-containing vesicles and cell death. *Biochimie* 94, 1262–1273. doi: 10.1016/j.biochi.2012.02.004

Mata, M., Delstanche, M., and Robert-Baudouy, J. (1978). Isolation of specialized transducing bacteriophages carrying the structural genes of the hexuronate system in *Escherichia coli* K-12: exu region. *J. Bacteriol.* 133, 549–557.

Maughan, H., Wilson, L. A., and Redfield, R. J. (2010). Bacterial DNA uptake sequences can accumulate by molecular drive alone. *Genetics* 186, 613–627. doi: 10.1534/genetics.110.119438

McDonough, M. A., and Butterton, J. R. (1999). Spontaneous tandem amplification and deletion of the Shiga toxin operon in *Shigella dysenteriae* 1. *Mol. Microbiol.* 34, 1058–1069. doi: 10.1046/j.1365-2958.1999.01669.x

McGrath, S., Fitzgerald, G. F., and Van Sinderen, D. (2002). Identification and characterization of phage-resistance genes in temperate lactococcal bacteriophages. *Mol. Microbiol.* 43, 509–520. doi: 10.1046/j.1365-2958.2002.02763.x

Mehta, P., Casjens, S., and Krishnaswamy, S. (2004). Analysis of the lambdoid prophage element e14 in the *E. coli* K-12 genome. *BMC Microbiol.* 4:4. doi: 10.1186/1471-2180-4-4

Mell, J. C., and Redfield, R. J. (2014). Natural competence and the evolution of DNA uptake specificity. *J. Bacteriol.* 196, 1471–1483. doi: 10.1128/JB.01293-13

Menouni, R., Hutinet, G., Petit, M. A., and Ansaldi, M. (2015). Bacterial genome remodeling through bacteriophage recombination. *FEMS Microbiol. Lett.* 362, 1–10. doi: 10.1093/femsle/fnu022

Mirold, S., Rabsch, W., Tschape, H., and Hardt, W. D. (2001). Transfer of the *Salmonella* type III effector sopE between unrelated phage families. *J. Mol. Biol.* 312, 7–16. doi: 10.1006/jmbi.2001.4950

Moore, R. A., Reckseidler-Zenteno, S., Kim, H., Nierman, W., Yu, Y., Tuanyok, A., et al. (2004). Contribution of gene loss to the pathogenic evolution of *Burkholderia pseudomallei* and *Burkholderia mallei*. *Infect. Immun.* 72, 4172–4187. doi: 10.1128/IAI.72.7.4172-4187.2004

Nadeem, A., and Wahl, L. M. (2017). Prophage as a genetic reservoir: promoting diversity and driving innovation in the host community. *Evolution* 71, 2080–2089. doi: 10.1111/evo.13287

Nasir, A., Kim, K. M., and Caetano-Anolles, G. (2017). Long-term evolution of viruses: a Janus-faced balance. *Bioessays* 39:1700026. doi: 10.1002/bies.201700026

Nilsson, A. S., Karlsson, J. L., and Haggard-Ljungquist, E. (2004). Site-specific recombination links the evolution of P2-like coliphages and pathogenic enterobacteria. *Mol. Biol. Evol.* 21, 1–13. doi: 10.1093/molbev/msg223

Obeng, N., Pratama, A. A., and Elsas, J. D. V. (2016). The significance of mutualistic phages for bacterial ecology and evolution. *Trends Microbiol.* 24, 440–449. doi: 10.1016/j.tim.2015.12.009

Ogier, J. C., Calteau, A., Forst, S., Goodrich-Blair, H., Roche, D., Rouy, Z., et al. (2010). Units of plasticity in bacterial genomes: new insight from the comparative genomics of two bacteria interacting with invertebrates, *Photorhabdus* and *Xenorhabdus. BMC Genomics* 11:568. doi: 10.1186/1471-2164-11-568

Osborne, S. E., Walthers, D., Tomljenovic, A. M., Mulder, D. T., Silphaduang, U., Duong, N., et al. (2009). Pathogenic adaptation of intracellular bacteria by rewiring a cis-regulatory input function. *Proc. Natl. Acad. Sci. U.S.A.* 106, 3982–3987. doi: 10.1073/pnas.0811669106

Palchevskiy, V., and Finkel, S. E. (2006). *Escherichia coli* competence gene homologs are essential for competitive fitness and the use of DNA as a nutrient. *J. Bacteriol.* 188, 3902–3910. doi: 10.1128/JB.01974-05

Papagiannis, C. V., Sam, M. D., Abbani, M. A., Yoo, D., Cascio, D., Clubb, R. T., et al. (2007). Fis targets assembly of the Xis nucleoprotein filament to promote excisive recombination by phage lambda. *J. Mol. Biol.* 367, 328–343. doi: 10.1016/j.jmb.2006.12.071

Parma, D. H., Snyder, M., Sobolevski, S., Nawroz, M., Brody, E., and Gold, L. (1992). The rex system of *Bacteriophage lambda*: tolerance and altruistic cell death. *Genes Dev.* 6, 497–510. doi: 10.1101/gad.6.3.497

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A. J., Thomson, N. R., et al. (2010). Antagonistic coevolution accelerates molecular evolution. *Nature* 464, 275–278. doi: 10.1038/nature08798

Peeters, S. H., and de Jonge, M. I. (2018). For the greater good: programmed cell death in bacterial communities. *Microbiol. Res.* 207, 161–169. doi: 10.1016/j.micres.2017.11.016

Picozzi, C., Meissner, D., Chierici, M., Ehrmann, M. A., Vigentini, I., Foschino, R., et al. (2015). Phage-mediated transfer of a dextranase gene in *Lactobacillus sanfranciscensis* and characterization of the enzyme. *Int. J. Food Microbiol.* 202, 48–53. doi: 10.1016/j.ijfoodmicro.2015.02.018

Ramisetty, B. C., Natarajan, B., and Santhosh, R. S. (2015). mazEF-mediated programmed cell death in bacteria: "what is this?". *Crit. Rev. Microbiol.* 41, 89–100. doi: 10.3109/1040841X.2013.804030

Ramisetty, B. C., and Santhosh, R. S. (2016). Horizontal gene transfer of chromosomal Type II toxin-antitoxin systems of *Escherichia coli*. *FEMS Microbiol. Lett.* 363:fnv238. doi: 10.1093/femsle/fnv238

Rangel, A., Steenbergen, S. M., and Vimr, E. R. (2016). Unexpected diversity of *Escherichia coli* sialate O-acetyl esterase NanS. *J. Bacteriol.* 198, 2803–2809. doi: 10.1128/JB.00189-16

Roossinck, M. J. (2011). The good viruses: viral mutualistic symbioses. *Nat. Rev. Microbiol.* 9, 99–108. doi: 10.1038/nrmicro2491

Ross, W., Landy, A., Kikuchi, Y., and Nash, H. (1979). Interaction of int protein with specific sites on lambda att DNA. *Cell* 18, 297–307. doi: 10.1016/0092-8674(79)90049-7

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. (2015a). VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. doi: 10.7717/peerj.985

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015b). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* 4:e08490. doi: 10.7554/eLife.08490

Saile, N., Voigt, A., Kessler, S., Stressler, T., Klumpp, J., Fischer, L., et al. (2016). *Escherichia coli* O157:H7 strain EDL933 harbors multiple functional prophage-associated genes necessary for the utilization of 5-N-Acetyl-9-O-Acetyl neuraminic acid as a growth substrate. *Appl. Environ. Microbiol.* 82, 5940–5950. doi: 10.1128/AEM.01671-16

Scanlan, P. D. (2017). Bacteria-bacteriophage coevolution in the human gut: implications for microbial diversity and functionality. *Trends Microbiol.* 25, 614–623. doi: 10.1016/j.tim.2017.02.012

Schindler, D., and Echols, H. (1981). Retroregulation of the int gene of *Bacteriophage lambda*: control of translation completion. *Proc. Natl. Acad. Sci. U.S.A.* 78, 4475–4479. doi: 10.1073/pnas.78.7.4475

Schmucker, R., Ritthaler, W., Stern, B., and Kamp, D. (1986). DNA inversion in bacteriophage Mu: characterization of the inversion site. *J. Gen. Virol.* 67(Pt 6), 1123–1133. doi: 10.1099/0022-1317-67-6-1123

Sechaud, L., Cluzel, P. J., Rousseau, M., Baumgartner, A., and Accolas, J. P. (1988). Bacteriophages of lactobacilli. *Biochimie* 70, 401–410. doi: 10.1016/0300-9084(88)90214-3

Sekulovic, O., and Fortier, L. C. (2015). Global transcriptional response of *Clostridium difficile* carrying the CD38 prophage. *Appl. Environ. Microbiol.* 81, 1364–1374. doi: 10.1128/AEM.03656-14

Shimada, K., Weisberg, R. A., and Gottesman, M. E. (1975). Prophage lambda at unusual chromosomal locations. III. The components of the secondary attachment sites. *J. Mol. Biol.* 93, 415–429. doi: 10.1016/0022-2836(75)90237-5

Shimizu, T., Ohta, Y., and Noda, M. (2009). Shiga toxin 2 is specifically released from bacterial cells by two different mechanisms. *Infect. Immun.* 77, 2813–2823. doi: 10.1128/IAI.00060-09

Silverman, M., and Simon, M. (1980). Phase variation: genetic analysis of switching mutants. *Cell* 19, 845–854. doi: 10.1016/0092-8674(80)90075-6

Sinha, S., and Redfield, R. J. (2012). Natural DNA uptake by *Escherichia coli*. *PLoS One* 7:e35620. doi: 10.1371/journal.pone.0035620

Song, Z., and Luo, L. (2012). *Escherichia coli* mutants induced by multi-ion irradiation. *J. Radiat. Res.* 53, 854–859. doi: 10.1093/jrr/rrs061

Soupene, E., Van Heeswijk, W. C., Plumbridge, J., Stewart, V., Bertenthal, D., Lee, H., et al. (2003). Physiological studies of *Escherichia coli* strain MG1655: growth defects and apparent cross-regulation of gene expression. *J. Bacteriol.* 185, 5611–5626. doi: 10.1128/JB.185.18.5611-5626.2003

Sousa, A. L. D., Maués, D., Lobato, A., Franco, E. F., Pinheiro, K., Araújo, F., et al. (2018). PhageWeb – Web interface for rapid identification and characterization of prophages in bacterial genomes. *Front. Genet.* 9:44. doi: 10.3389/fgene.2018.00644

Steinberg, K. M., and Levin, B. R. (2007). Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proc. Biol. Sci.* 274, 1921–1929. doi: 10.1098/rspb.2007.0245

Susskind, M. M., Wright, A., and Botstein, D. (1974). Superinfection exclusion by P22 prophage in lysogens of *Salmonella typhimurium*. IV. Genetics and physiology of sieB exclusion. *Virology* 62, 367–384. doi: 10.1016/0042-6822(74)90399-7

Svab, D., Horvath, B., Maroti, G., Dobrindt, U., and Toth, I. (2013). Sequence variability of P2-like prophage genomes carrying the cytolethal distending toxin V operon in *Escherichia coli* O157. *Appl. Environ. Microbiol.* 79, 4958–4964. doi: 10.1128/AEM.01134-13

Swalla, B. M., Cho, E. H., Gumport, R. I., and Gardner, J. F. (2003). The molecular basis of co-operative DNA binding between lambda integrase and excisionase. *Mol. Microbiol.* 50, 89–99. doi: 10.1046/j.1365-2958.2003.03687.x

Tan, Q., Awano, N., and Inouye, M. (2011). YeeV is an *Escherichia coli* toxin that inhibits cell division by targeting the cytoskeleton proteins, FtsZ and MreB. *Mol. Microbiol.* 79, 109–118. doi: 10.1111/j.1365-2958.2010.07433.x

Tao, L., Wu, X., and Sun, B. (2010). Alternative sigma factor sigmaH modulates prophage integration and excision in *Staphylococcus aureus*. *PLoS Pathog.* 6:e1000888. doi: 10.1371/journal.ppat.1000888

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702. doi: 10.1038/nrg3053

Thompson, J. F., Waechter-Brulla, D., Gumport, R. I., Gardner, J. F., Moitoso, De Vargas, L., et al. (1986). Mutations in an integration host factor-binding site: effect on lambda site-specific recombination and regulatory implications. *J. Bacteriol.* 168, 1343–1351. doi: 10.1128/jb.168.3.1343-1351.1986

Toba, F. A., Thompson, M. G., Campbell, B. R., Junker, L. M., Rueggeberg, K. G., and Hay, A. G. (2011). Role of DLP12 lysis genes in *Escherichia coli* biofilm formation. *Microbiology* 157, 1640–1650. doi: 10.1099/mic.0.045161-0

Varga, M., Pantucek, R., Ruzickova, V., and Doskar, J. (2016). Molecular characterization of a new efficiently transducing bacteriophage identified in meticillin-resistant *Staphylococcus aureus*. *J. Gen. Virol.* 97, 258–268. doi: 10.1099/jgv.0.000329

Veening, J. W., and Blokesch, M. (2017). Interbacterial predation as a strategy for DNA acquisition in naturally competent bacteria. *Nat. Rev. Microbiol.* 15, 621–629. doi: 10.1038/nrmicro.2017.66

Ventura, M., Canchaya, C., Kleerebezem, M., De Vos, W. M., Siezen, R. J., and Brussow, H. (2003). The prophage sequences of *Lactobacillus plantarum* strain WCFS1. *Virology* 316, 245–255. doi: 10.1016/j.virol.2003.08.019

Veses-Garcia, M., Liu, X., Rigden, D. J., Kenny, J. G., Mccarthy, A. J., and Allison, H. E. (2015). Transcriptomic analysis of Shiga-toxigenic bacteriophage carriage reveals a profound regulatory effect on acid resistance in *Escherichia coli*. *Appl. Environ. Microbiol.* 81, 8118–8125. doi: 10.1128/AEM.02034-15

von Wintersdorff, C. J., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., et al. (2016). Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* 7:173. doi: 10.3389/fmicb.2016.00173

Waldor, M. K., and Mekalanos, J. J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 272, 1910–1914. doi: 10.1126/science.272.5270.1910

Wang, P. W., Chu, L., and Guttman, D. S. (2004). Complete sequence and evolutionary genomic analysis of the *Pseudomonas aeruginosa* transposable bacteriophage D3112. *J. Bacteriol.* 186, 400–410. doi: 10.1128/JB.186.2.400-410.2004

Wang, X., Kim, Y., Ma, Q., Hong, S. H., Pokusaeva, K., Sturino, J. M., et al. (2010). Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* 1:147. doi: 10.1038/ncomms1146

Wang, X., and Wood, T. K. (2016). Cryptic prophages as targets for drug development. *Drug Resist. Updates* 27, 30–38. doi: 10.1016/j.drup.2016.06.001

Wipf, J. R., Schwendener, S., and Perreten, V. (2014). The novel macrolide-lincosamide-streptogramin B resistance gene erm(44) is associated with a prophage in *Staphylococcus xylosus*. *Antimicrob. Agents Chemother.* 58, 6133–6138. doi: 10.1128/AAC.02949-14

Yin, S., Bushman, W., and Landy, A. (1985). Interaction of the lambda site-specific recombination protein Xis with attachment site DNA. *Proc. Natl. Acad. Sci. U.S.A.* 82, 1040–1044. doi: 10.1073/pnas.82.4.1040

Zhang, L., Zhu, B., Dai, R., Zhao, G., and Ding, X. (2013). Control of directionality in *Streptomyces phage* phiBT1 integrase-mediated site-specific recombination. *PLoS One* 8:e80434. doi: 10.1371/journal.pone.0080434

Zschöck, M., Botzler, D., Blöcher, S., Sommerhäuser, J., and Hamann, H. P. (2000). Detection of genes for enterotoxins (ent) and toxic shock syndrome toxin-1 (tst) in mammary isolates of *Staphylococcus aureus* by polymerase-chain-reaction. *Int. Dairy J.* 10, 569–574. doi: 10.1016/S0958-6946(00)00084-4

Conflict of Interest Statement: 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.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 08 March 2019
doi: 10.3389/fmicb.2019.00430

Pandoravirus Celtis Illustrates the Microevolution Processes at Work in the Giant *Pandoraviridae* Genomes

Matthieu Legendre[1], Jean-Marie Alempic[1], Nadège Philippe[1], Audrey Lartigue[1], Sandra Jeudy[1], Olivier Poirot[1], Ngan Thi Ta[1], Sébastien Nin[1], Yohann Couté[2], Chantal Abergel[1]* and Jean-Michel Claverie[1]*

[1] Aix Marseille Univ, CNRS, IGS, Structural and Genomic Information Laboratory (UMR7256), Mediterranean Institute of Microbiology (FR3479), Marseille, France, [2] Inserm, BIG-BGE, CEA, Université Grenoble Alpes, Grenoble, France

With genomes of up to 2.7 Mb propagated in μm-long oblong particles and initially predicted to encode more than 2000 proteins, members of the *Pandoraviridae* family display the most extreme features of the known viral world. The mere existence of such giant viruses raises fundamental questions about their origin and the processes governing their evolution. A previous analysis of six newly available isolates, independently confirmed by a study including three others, established that the *Pandoraviridae* pan-genome is open, meaning that each new strain exhibits protein-coding genes not previously identified in other family members. With an average increment of about 60 proteins, the gene repertoire shows no sign of reaching a limit and remains largely coding for proteins without recognizable homologs in other viruses or cells (ORFans). To explain these results, we proposed that most new protein-coding genes were created *de novo*, from pre-existing non-coding regions of the G+C rich pandoravirus genomes. The comparison of the gene content of a new isolate, pandoravirus celtis, closely related (96% identical genome) to the previously described p. quercus is now used to test this hypothesis by studying genomic changes in a microevolution range. Our results confirm that the differences between these two similar gene contents mostly consist of protein-coding genes without known homologs, with statistical signatures close to that of intergenic regions. These newborn proteins are under slight negative selection, perhaps to maintain stable folds and prevent protein aggregation pending the eventual emergence of fitness-increasing functions. Our study also unraveled several insertion events mediated by a transposase of the hAT family, 3 copies of which are found in p. celtis and are presumably active. Members of the *Pandoraviridae* are presently the first viruses known to encode this type of transposase.

Keywords: *de novo* gene creation, comparative genomics, *Acanthamoeba*, giant viruses, soil viruses, hAT transposase

OPEN ACCESS

Edited by:
Ricardo Flores,
Instituto de Biología Molecular y Celular de Plantas (IBMCP), Spain

Reviewed by:
Amparo Latorre,
University of Valencia, Spain
Victor Krylov,
I. I. Mechnikov Research Institute for Vaccines and Sera (RAS), Russia

****Correspondence:***
Chantal Abergel
chantal.abergel@igs.cnrs-mrs.fr
Jean-Michel Claverie
jean-michel.claverie@univ-amu.fr

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

Received: *18 December 2018*
Accepted: *19 February 2019*
Published: *08 March 2019*

Citation:
Legendre M, Alempic J-M, Philippe N, Lartigue A, Jeudy S, Poirot O, Ta NT, Nin S, Couté Y, Abergel C and Claverie J-M (2019) Pandoravirus Celtis Illustrates the Microevolution Processes at Work in the Giant Pandoraviridae Genomes. Front. Microbiol. 10:430. doi: 10.3389/fmicb.2019.00430

INTRODUCTION

The *Pandoraviridae* is a proposed family of giant dsDNA viruses - not yet registered by the International Committee on Taxonomy of Viruses (ICTV) - multiplying in various species of Acanthamoeba through a lytic infectious cycle. Their linear genomes, flanked by large terminal repeats, range from 1.9 to 2.7 Mb in size, and are propagated in elongated oblong particles approximately 1.2 μm long and 0.6 μm in diameter (**Supplementary Figure S1**). The prototype strain (and the one with the largest known genome) is pandoravirus salinus (p. salinus), isolated from shallow marine sediments off the coast of central Chile (Philippe et al., 2013). Other members were soon after isolated from worldwide locations. Complete genome sequences have been determined for p. dulcis (Melbourne, Australia) (Philippe et al., 2013), p. inopinatum (Germany) (Antwerpen et al., 2015), p. macleodensis (Melbourne, Australia), p. neocaledonia (New Caledonia), and p. quercus (France) (Legendre et al., 2018), and three isolates from Brazil (p. braziliensis, p. pampulha, and p. massiliensis) (Aherfi et al., 2018). A standard phylogenetic analysis of the above strains suggested that the *Pandoraviridae* family consists of two separate clades (Claverie et al., 2018; **Figure 1**). The average proportion of identical amino acids between pandoravirus orthologs within each clade is above 70% while it is below 55% between members of the A and B clades. Following a stringent reannotation of the predicted protein-coding genes using transcriptomic and proteomic data, our comparative genomic analysis reached the main following conclusions (Legendre et al., 2018):

(1) The uniquely large proportion of predicted proteins without homologs outside of the *Pandoraviridae* (ORFans) is real and not due to bioinformatic errors induced by the above-average G+C content (>60%) of pandoravirus genomes;
(2) the *Pandoraviridae* pan genome appears "open" (i.e., unbounded);
(3) as most of the genes are unique to each strain are ORFans, they were not horizontally acquired from other (known) organisms;
(4) they are neither predominantly the result of gene duplications.

The scenario of *de novo* and *in situ* gene creation, supported by the analysis of their sequence statistical signatures, thus became our preferred explanation for the origin of strain specific genes.

In the present study, we take advantage of the high similarity (96.7% DNA sequence identity) between p. quercus and a newly characterized isolate, p. celtis, to investigate the microevolution processes initiating the divergence between pandoraviruses. Our results further support *de novo* gene creation as a main diversifying force of the *Pandoraviridae* family.

MATERIALS AND METHODS

Virus Isolation, Production, and Purification

P. celtis and p. quercus were isolated in November 2014 from samples of surface soil taken at the base of two trees (*Celtis australis* and *Quercus ilex*) less than 50 m apart in an urban green space of Marseille city (GPS: 43°15′16.00′′N, 5°25′4.00′′E). Their particles were morphologically identical to previously characterized pandoraviruses (**Supplementary Figure S1**). The viral populations were amplified by co-cultivation with *Acanthamoeba castellanii*. They were then cloned, mass-produced and purified as previously described (Philippe et al., 2013).

Genome and Transcriptome Sequencing, Annotation

The p. celtis genome was fully assembled from one PacBio SMRT cell sequence data with the HGAP 4 assembler (Chin et al., 2013) from the SMRT link package version 5.0.1.9585 with default parameters and the "aggressive" option = true. Genome polishing was finally performed using the SMRT package. Stringent gene annotation was performed as previously described (Legendre et al., 2018). Briefly, data from proteomic characterization of the purified virions were combined with stranded RNA-seq transcriptomic data, as well as protein homology among previously characterized pandoraviruses. The transcriptomic data were generated from cells collected every hour over an infectious cycle of 15 h. They were pooled and RNAs were extracted prior poly(A)$^+$ enrichment. The RNA were then sent for sequencing. Stranded RNA-seq reads were then used to accurately annotate protein-coding as well as non-coding RNA genes. A threshold of gene expression (median read coverage > 5 over the whole transcript) larger than the lowest one associated to proteins detected in proteomic analyses was required to validate all predicted genes (including novel genes) (**Table 1**). Genomic regions exhibiting similar expression levels but did not encompass predicted proteins or did overlap with protein-coding genes expressed from opposite strands were annotated as "non-coding RNA" (ncRNA) after assembly by Trinity (Grabherr et al., 2011). When only genomic data were available, namely for p. pampulha, p. massiliensis, and p. braziliensis (Aherfi et al., 2018), we annotated protein-coding genes using *ab initio* prediction coupled with sequence conservation information as previously described (Legendre et al., 2018).

Gene clustering was performed on all available pandoraviruses' protein sequences using Orthofinder (Emms and Kelly, 2015) with defaults parameters except for the "msa" option for the gene tree inference method.

Functional annotation of protein-coding genes was performed using a combination of protein domains search with the CD-search tool (Marchler-Bauer and Bryant, 2004) and HMM-HMM search against the Uniclust30 database with the HHblits tool (Remmert et al., 2011). In addition, we used the same procedure to update the functional gene annotation of p. salinus, p. dulcis, p. quercus, p. macleodensis, and p. neocaledonia (Genbank IDs: KC977571, KC977570, MG011689, MG011691, and MG011690).

Phylogenetic and Selection Pressure Analysis

The phylogenetic tree (**Figure 1**) was computed from the concatenated multiple alignment of the sequences of pandoraviruses' core proteins corresponding to single-copy genes. The alignments of orthologous genes peptide sequences were done using Mafft (Katoh et al., 2002). The tree was computed using IQtree (Hoang et al., 2018) with the following options: -m MFP –bb 10000 –st codon –bnni. The best model chosen was: GY+F+R5. Codon sequences were subsequently

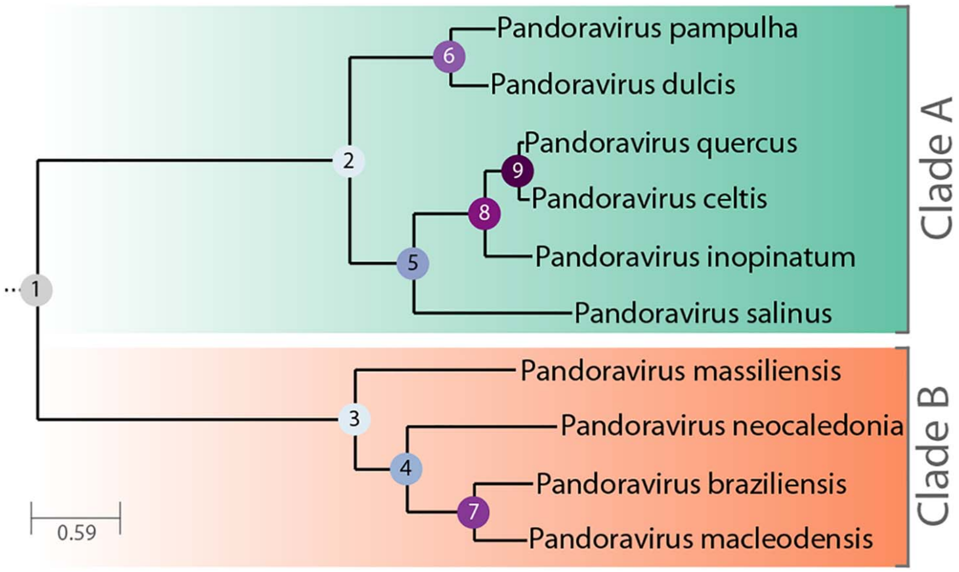


FIGURE 1 | Phylogenetic tree of known, fully characterized pandoraviruses. The tree was computed from the pandoraviruses' core genes (see Materials and Methods for details). The estimated bootstrap values were all equal to 1 and thus not reported. Internal nodes are labeled according to their depth in the tree.

mapped on these alignments. Ratios of non-synonymous (dN) over synonymous (dS) mutation rates for pairs of orthologous genes were computed using the YN00 method from the PAML package (Yang, 2007). Filters were applied so that dN/dS ratios were only considered if: $dN > 0$, $dS > 0$, $dS \leq 2$ and $dN/dS \leq 10$. We also computed the Codon Adaptation Index (CAI) of p. celtis genes using the cai tool from the Emboss package (Rice et al., 2000) as previously described (Legendre et al., 2018).

Particle Proteomics

The p. celtis particle proteome was characterized by mass spectrometry-based proteomics from purified viral particles as previously described (Legendre et al., 2015, 2018).

RESULTS

Main Structural Features of the P. celtis and P. quercus Genomes

The p. celtis dsDNA genome sequence was assembled as a single 2,028,440 bp-long linear contig, thus slightly shorter than the published 2,077,288-bp for p. quercus. Both contains 61% of G+C. The two genomes exhibit a global collinearity well illustrated by a dotplot comparison of their highly similar nucleotide sequences (**Figure 2**). In particular, their 48.8 kb difference in genome size does not obviously correspond to a large non-homologous region or a large-scale duplication. Both genomes begin by a nearly perfect 19-kb long palindrome (labeled "P" in **Figure 2**). As we did not observe this feature in the other published pandoravirus sequences, it may be specific of p. celtis and p. quercus, or its absence in other genomes may result from flaws in the assembly of terminal sequences due to insufficient read coverage or quality. Six kb downstream, p. quercus exhibits a segment ([25,480–43,420]) nearly identical (20 indels) to the distal end of the genome, inverted (labeled "T" in **Figure 2**). Remnants of a similar feature appear blurred in p. celtis, but are absent from the other pandoravirus genomes. As these regions are accurately determined, we can infer that a duplication followed by an inversion/translocation of the distal genome terminus occurred in the ancestor of p. quercus and p. celtis (between node 8 and 9 in **Figure 1**).

The next noticeable structural rearrangements consist of three segments denoted S_0, S_1, and S_2 in **Figure 2**. Each of these segments are flanked by terminal inverted repeats (**Supplementary Figure S2**) and encode a protein (respectively pclt_cds_98, pclt_cds_672, and pclt_cds_871) exhibiting both a BED zinc finger domain and a C-terminal dimerisation region, typical of transposases of the hobo/Activator/Tam3 (hAT) family. All these proteins exhibit the intact signature of hAT transposases and are thus probably active (Atkinson, 2015). Using these sequences as queries, we readily identified other well-conserved hAT transposase homologs in p. pampulha (ppam_cds_67, cds_531, cds_663), p. macleodensis (pmac_cds_424, cds_799, cds 869) and p. neocaledonia (pneo_cds_387, cds_113, cds_658, cds_798).

Besides their common transposase, the S_0 [10.3 kb, pos. 152,985–163,291], S_1 [10.3 kb, pos. 1,156,443–1,166,965], and S_2 [7.3 kb, pos. 1,452,773–1,460,053] transposons encode different sets of proteins. S_0 encodes 9 proteins (pclt_cds_90-98). Except for the transposase, all of them have no predicted function, and no recognizable homologs outside of the *Pandoraviridae* (i.e., they are "family ORFans"). S_1 encodes 11 proteins (pclt_cds_672-682) all of which also have no functional attributes and are family ORFans, except for the transposase. S_2 encodes 7 proteins (pclt_cds_865-871), all of which have no recognizable signature (except for the transposase and a F-box domain for

TABLE 1 | P. celtis and p. quercus unique protein-coding genes.

Gene #	Size (aa)	Most ancestral detection	Homolog in p. quercus	Predicted DNA binding	Median RNAseq read coverage
pclt_cds_11	69	Since node 5	None	Yes	704
pclt_cds_308	181	Since node 2	pqer_ncRNA_47	No	986
pclt_cds_350	121	Since node 9	Intergenic	Yes	57
pclt_cds_376	145	Since node 5	3′UTR pqer_cds_371	Yes	454
pclt_cds_725	104	None	None	Yes	46
pclt_cds_870	205	Since node 2	None	Yes	2027
pclt_cds_995	149	Since node 5	5′UTR pqer_cds_981	Yes	13
pclt_cds_1081	125	Since node 8	5′UTR pqer_cds_1061	Yes	12,090
pclt_cds_1084	114	Since node 8	intergenic	Yes	130
			Homolog in p. celtis		
pqer_cds_6	76	Since node 8	None	Yes	311
pqer_cds_13	117	Since node 5	Intergenic	Yes	99
pqer_cds_17	82	Since node 2	Intergenic	Yes	482
pqer_cds_53	85	Since node 8	Intergenic	Yes	48
pqer_cds_143	93	Since node 8	Anti 5′UTR pclt_cds_146	Yes	177
pqer_cds_151	71	Since node 9	Intergenic	Yes	21
pqer_cds_203	101	Since node 8	Antisense pclt_cds_206	Yes	528
pqer_cds_350	94	None	None	No	160
pqer_cds_474	74	Since node 9	Intergenic	Yes	114
pqer_cds_486	146	Since node 2	3′UTR pclt_cds_499	Yes	71
pqer_cds_665	114	None	None	Yes	1,955
pqer_cds_673	383	None	None	Yes	656
pqer_cds_685	124	Since node 8	Alternative frame pclt_cds_685	No	117
pqer_cds_736	121	Since node 9	Intergenic	Yes	70
pqer_cds_875	136	None	None	No	2,050
pqer_cds_876	74	Since node 2	None	No	1,695
pqer_cds_877	78	None	None	Yes	320
pqer_cds_878	224	None	None	Yes	231
pqer_cds_1061	152	Since node 8	Alternative frame pclt_cds_1081	Yes	14,186
pqer_cds_1178	84	Since node 9	Anti 5′UTR pclt_cds_1203	Yes	170
pqer_cds_1183	158	Since node 2	None	Yes	198

pclt_cds_870). Besides the transposase, a single protein have paralogs in the S_0, S_1, and S_2 transposons (pclt_cds_92, 681, 869), and one is only shared by S_0 and S_1 (pclt_cds_94, 678). These differences clearly suggest that S_0, S_1, S_2 are not the results of recent duplication/transposition events from a common template.

A tentative scenario for the insertion of the p. celtis hAT transposons was inferred from their presence/absence in p. quercus and the sequence similarity of the transposases. The S_1 transposon is the only one shared between the two strains (**Supplementary Figure S2**). Moreover, all the orthologous proteins encoded in S_1 are 100% identical (pclt_cds_672-682 vs. pqer_cds_619-610), including the transposases (pclt_cds_672 and pqer_cds_619). A first possibility is that the S_1 segment was already present in the ancestor of p. celtis and p. quercus (**Figure 1**, node 9), then was inverted and translocated about 70 kb downstream from its initial location in p. celtis. However, since this transposon is absent from p. inopinatum (diverging after node 8, **Figure 1**), it may have been independently gained from the same source into p. celtis and p. quercus just after their divergence as two variants within the local viral population. Interestingly, an unrelated sequence of 30.5 kb was inserted at the homologous positions in p. quercus (pos. 1,177,379–1,207,935) (**Supplementary Figure S2**). This insertion encodes 24 proteins (pqer_cds_659-682) 13 of which are anonymous and ORFans (i.e., only homologous to other pandoraviruses' proteins), the other exhibiting uninformative motifs such as ankyrin repeats (pqer_cds_668, 669, 674-676, 680), F-box domains (pqer_cds_681, 682), Morn repeat (pqer_cds_661) and Ring domain (pqer_cds_670). One protein (pqer_cds_673) exhibits a low ($E < 10^{-2}$) and partial similarity with a domain found at the N terminus of structural maintenance of chromosomes (SMC) proteins. However, none of the proteins encoded by this insertion bears any similarity with a hAT family transposase making the mechanism and the origin of this insertion all the more puzzling.

Finally, the proposed scenario concerning S_0 and S_2 are simpler. As the two corresponding transposases (pclt_cds_98 and pclt_cds_871) are only 91% identical to each other and less than 67% identical to pclt_cds_672, we propose that they resulted from two independent insertion events that occurred after the

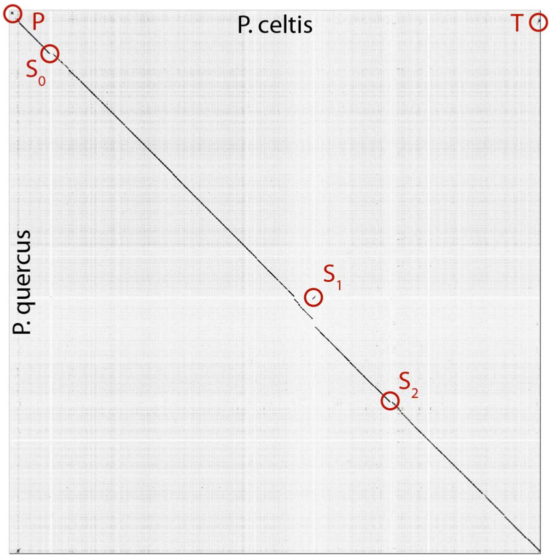


FIGURE 2 | Dot plot (nucleotide) comparison of the p. celtis vs. p. quercus genomes. Computations and display were generated using Gepard (Krumsiek et al., 2007). Specific regions, labeled P, T, S_0, S_1, and S_2 are described in the main text.

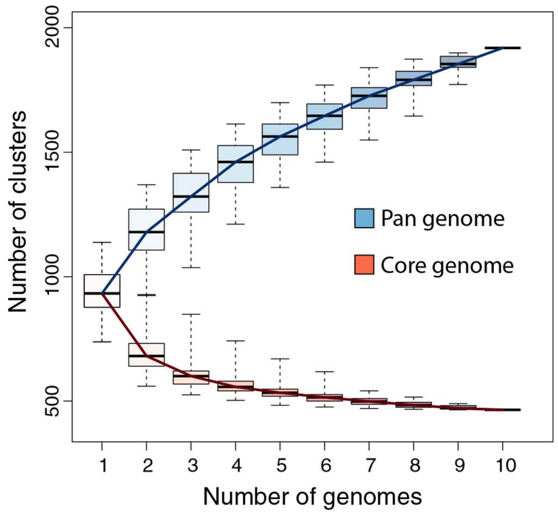


FIGURE 3 | *Pandoraviridae* pan-genome and core-genome. The boxplot represents the number of protein clusters as a function of the number of sequenced genomes in all possible combinations. The whiskers correspond to the extreme data points.

p. quercus/p. celtis divergence, from distinct templates with little overlap in their gene cargo (pclt_cds_92 and 869) besides active transposases. An alternative scenario, would assume the presence of S_0 and S_2 in the p. quercus/p. celtis ancestor, and their subsequent excision in p. quercus. This scenario is less likely as p. inopinatum (**Figure 1**, node 8) does not encode any hAT transposase.

Tandem Duplications

Approximately 28 kb downstream from the S_2 transposon (absent in p. quercus), the p. celtis *vs.* p. quercus dot plot indicates a cluster of closely interspersed direct repeats (**Supplementary Figure S3**) coding for paralogous proteins that all contain a highly conserved N-terminal fascin-like domain (CD00257). This 80-residue long domain is found in actin-bundling/crosslinking proteins. Many copies of these proteins are found in each of the known pandoraviruses. There are 17 (full-length) copies in p. quercus, and 14 in p. celtis, the paralogs sharing from 100 to 40% identical residues. Using standard phylogenetic analysis each of the p. celtis paralogs are unambiguously clustered with p. quercus counterparts in 14 orthologous pairs, indicating that the multiplication of these proteins took place prior to their divergence (**Supplementary Figure S3**). As the three additional p. quercus paralogs (pqer_cds_130, 131, and 870) with no obvious counterpart in p. celtis are not very similar to other p. quercus paralogs (max is 72% sequence identity between pqer_cds_870 and pqer_cds_869), it is unlikely that they arise from post-divergence duplications. The most parsimonious scenario is thus that their counterparts have been recently lost by p. celtis. In a dot plot, the pqer_cds_130 and pqer_cds_131 genes correspond to a 4 kb deletion at position 224,350 in p. celtis. The remaining pqer_cds_870 is located in a dense cluster of 6 contiguous paralogs (pqer_cds_866-871) in p. quercus, clearly homologous to a similar cluster of 5 contiguous paralogs (pclt_cds_884-888) in p. celtis (**Supplementary Figure S3**). The deletion of a short segment of the p. celtis genome is visible at this exact position (1,487,300). Another member of this cluster (pclt_cds_886) just entered a pseudogenization process through the several insertion/deletions breaking its original reading frame (**Supplementary Figure S3**).

Core Genes and Pan Genome Update

We previously designed a stringent gene annotation scheme to minimize overpredictions in G+C–rich genomes. In particular, genes predicted to encode proteins without database homolog (i.e., ORFans) were required to overlap with detected sense transcripts to be validated (Legendre et al., 2018). According to this protocol, p. celtis encodes at least 1184 protein-coding genes (CDS) compared to 1185 protein-coding genes for p. quercus. Both encode a single tRNA(Pro). It is worth noticing that despite a strong reduction compared to the number of proteins initially predicted by standard methods (Philippe et al., 2013), the proportion of ORFans (i.e., w/o homolog outside of the *Pandoraviridae* family) remains high at 68%. This confirms that the lack of database homologs is not caused by bioinformatics overpredictions (Legendre et al., 2018). P. celtis and p. quercus shared an average of 96% identical residues, as computed from 822 unambiguous orthologous proteins encoded by single copy genes (i.e., w/o paralogs).

We previously showed that the comparison of the first six available pandoravirus genomes rapidly converged toward a relatively small set of common genes, corresponding to 455 distinct proteins (i.e., clusters) (Legendre et al., 2018). An even smaller estimate (352) was subsequently proposed by another laboratory using a different clustering protocol and strains (Aherfi et al., 2018). Now based on the ten fully sequenced pandoraviruses and an optimized clustering protocol (see Materials and Methods), the *Pandoraviridae* core gene set was found to contain 464 genes, close to the asymptotic limit suggested by **Figure 3**. Including p. celtis in the analysis caused the removal of two proteins from the list. Their homologs in p. quercus are pqer_cds_672, and pqer_cds_370, two proteins w/o recognizable functional attribute or homolog outside of the *Pandoraviridae*. Compared to the average number of distinct proteins encoded by individual pandoravirus genomes, the proportion of those presumably essential is thus less than half. On the other hand, the predicted gene content of p. celtis further increased the *Pandoraviridae* pan-genome. Despite its overall close similarity with p. quercus, p. celtis remains on a growth curve whereby each new randomly isolated pandoravirus is predicted to add about 60 distinct proteins to the total number of protein clusters already identified in the family (**Figure 3**). The process by which such addition could arise was further investigated by a detailed comparison of the genomes of the most similar relatives, p. quercus and p. celtis.

Protein-Coding Genes Unique to P. celtis or P. quercus

Given the strong similarity between the two genomes, determining their difference in gene content is straightforward. However, these differences can be due to various mechanisms: the duplication of existing genes, their differential loss, their acquisition by horizontal transfer, or their *de novo* creation. Here we wanted to focus on the later mechanism, previously proposed to be prominent in the *Pandoraviridae* family (Legendre et al., 2018). The candidate genes most likely resulting from *de novo* creation would be those unique to p. celtis or p. quercus. If there are no homolog in any of the other pandoraviruses, *de novo* creation (or acquisition from an unknown source) becomes the most parsimonious evolutionary scenario, compared to a scenario whereby such a gene, present in an ancestral pandoravirus would have been independently lost multiple times. The 30 protein-coding genes unique to p. celtis ($N = 9$) and p. quercus ($N = 21$) are listed in **Table 1**. As previously noticed (Legendre et al., 2018), these presumed novel genes are associated with statistics that are significantly different from other pandoravirus protein-coding genes. They exhibit a lower Codon Adaptation Index (CAI = 0.233 vs. 0.351, Wilcoxon test, $p < 10^{-14}$), a lower G+C content (57.5% vs. 64.4%, $p < 10^{-16}$), and encode smaller proteins [length (aa) = 126.6 vs. 387.3, $p < 10^{-15}$]. The two later characteristics make them similar to the random ORFs found in intergenic regions (i.e., the so-called "protogenes") from which we proposed they originated.

We also noticed that the proteins encoded by the novel genes exhibit an amino-acid composition widely different from the rest of the predicted proteome (Chi-2, df = 19, $p < 10^{-10}$), with the largest variations observed for lysine (increasing from 2 to 5%), phenylalanine (2.2 to 4%), and arginine (8.6 to 11.3%). Owing to the anomalous proportions of these residues, 25 out of the 30 novel p. celtis and p. quercus unique proteins are predicted to be DNA-binding (Szilágyi and Skolnick, 2006) (**Table 1**). Although such predictions are subject to caution (as the predictions remains the same for shuffled sequences) binding novel proteins to the viral genome might help segregate loosely folded (i.e., intrinsically disordered) proteins and diminish their potential interference with the rest of the viral proteome. Over time, these proteins may also gain some regulatory functions.

We investigated the putative intergenic origin of these 30 strain-specific protein-coding genes by searching for remote sequence similarity in the genome of all other pandoraviruses, using tblastn (protein query against six reading frame translation, $E < 10^{-3}$) (Sayers et al., 2019). For seven of them, one unique to p. celtis and six unique to p. quercus, we found no similarity within the non-coding moiety of other pandoraviruses. Our most parsimonious interpretation is that they represent recent independent additions in each genome by *de novo* creation or transfer from an unknown organism w/o known relative. A slightly less parsimonious scenario would be their addition prior to node 9, followed by differential losses in p. quercus or p. celtis. Remote but significant ($E < 10^{-3}$) similarities were detected for the 25 other novel genes within non-coding regions of other pandoraviruses, all of them members of Clade A (**Figure 1**). These positive matches were distributed as followed: five in strains that diverged since node 9, eight in all strains that diverged since node 8, four in all strains that diverged since node 5, and six in all strains that diverged since node 2. We did not detect significant matches in earlier diverging Clade B members. The distribution of these traces in pandoraviruses at various evolutionary distances is interpreted in the Section "Discussion."

ncRNAs

We annotated 161 ncRNAs in p. celtis (**Supplementary Table S1**), a number close to the 157 ncRNAs predicted for p. quercus using the same protocol. Such numerous transcripts without protein coding capability were previously noticed for other pandoravirus strains (Legendre et al., 2018). The p. celtis ncRNAs vary in length from 234 to 4,456 nucleotides (median = 1384, mean = 1413 ± 639). By comparison, the p. quercus ncRNAs range from 273 to 5,176 (median = 1,189, mean = 1299 ± 738). The two distributions are not significantly different (T-test: $p > 0.13$). The expression levels (in median read coverage) of p. celtis ncRNAs vary from 92 to 1608 (median = 283) compared to 3665–506,994 (median = 690) for protein coding transcripts. A similar non-coding *versus* coding expression median ratio (≈ 0.41) was previously found for p. quercus (ncRNAs median = 228, protein-coding transcript median = 551), in a distinct experiment. As previously noticed for other pandoravirus strains, most of p. celtis ncRNAs (154/161 = 95.6%) are overlapping by more than half of their length with a protein-coding transcript expressed from the opposite strand, and only 7 are mostly intergenic.

Although the p. celtis and p. quercus genomes share a very large proportion of their protein-coding genes (1146/1184 = 96.8%), this was not the case for ncRNAs. We found that only 87 of p. celtis ncRNAs (87/161 = 54%) exhibit a homolog among p. quercus ncRNAs (**Supplementary Table S1**). The 74 p. celtis ncRNAs without homologs thus mostly correspond to the lack of detectable transcription in the cognate sequence of the p. quercus genome. Unexpectedly, the p. celtis and p. quercus novel protein-coding genes discussed above (see Protein-Coding Genes Unique to p. celtis or p. quercus) only rarely overlap with ncRNAs in other strains. The single case is pclt_cds_308, the coding region of which overlaps with a p. quercus ncRNA (pqer_ncRNA_47) (**Table 1**).

Selection Pressure on New Genes

Protein-coding genes shared by at least two different pandoraviruses provide an opportunity to estimate the selection pressure acting on them by computing the ratio of nonsynonymous (dN) to synonymous (dS) substitutions per site. We computed the dN/dS ratio of genes shared by increasingly close pandoraviruses, dating their creation or acquisition by reference to the corresponding node in the pandoravirus phylogenetic tree (nodes 1–9 in **Figure 1**). All pairs of orthologous gene sequences were thus analyzed using the YN00 algorithm from the PAM package (Yang, 2007) and assigned to their most likely creation/acquisition node based on their presence or absence in various clades of pandoraviruses. As shown in **Figure 4** all genes are on average under negative selection pressure ($dN/dS < 1$), including the presumably most recently created/acquired genes as those only found in p. quercus and p. celtis (i.e., node 9 in **Figure 1**). As previously documented for short evolutionary distances and very close gene sequences (our case), the computed dN/dS values are probably overestimated (i.e., closer to 1) given that a fraction of the deleterious mutations might not yet be fixed (Rocha et al., 2006). This is consistent with the observed negative correlation between the depth of creation/acquisition and the computed purifying selection (**Figure 4**). In other words recently created/acquired genes appear under selective constraints weaker than that of "older" genes. The fact that dN/dS values tend to decrease with longer divergence times was previously noticed (Rocha et al., 2006).

DISCUSSION

Following our previous analysis of 6 available members of the *Pandoraviridae* family, we proposed that the gigantism of their genomes as well as the large proportion of ORFans among their encoded proteins was the result of unusual evolutionary mechanisms, including *de novo* gene creation from previously non-coding sequences (Legendre et al., 2018). In this previous study, we compared pandoravirus isolates exhibiting pairwise similarity ranging from 54 to 88%, as computed from a super alignment of their orthologous proteins. At such evolutionary distances, the observed differences are most often the result of multiple and overlapping elementary variation processes the succession of which becomes impossible to retrace. The subsequent isolation of p. celtis, a new pandoravirus strain very similar to p. quercus (with orthologous proteins sharing 96% of identical residues in average), gave us the opportunity of identifying and estimating the relative contributions of various types of genomic alterations at work during their microevolution from their recent common ancestor.

The global analyses of the gene contents of the 10 members of the *Pandoraviridae* family available today confirmed previous estimates of the number of core gene clusters (Aherfi et al., 2018; Legendre et al., 2018) at 464. Such a small proportion (less than half) of presumably "essential" genes compared to the total number of proteins encoded by each pandoravirus genome (ranging from 1070 to 1430, using our stringent protocol) (Legendre et al., 2018) raises the question of the origin and utility of so many "accessory" genes. Non-essential genes are normally eliminated from the genomes of obligate intracellular parasites or symbionts through reductive evolution (Corradi et al., 2010; Lopez-Madrigal et al., 2011; McCutcheon and Moran, 2011; Latorre and Manzano-Marín, 2017; Floriano et al., 2018). The contrast is even more puzzling when the size of the *Pandoraviridae* pan-genome shows no sign of leveling off after reaching 1910 different protein-coding gene clusters, sustaining a trend predicting that each new isolate will contribute 60 additional clusters. Understanding the mechanism by which new genes, - most of which encode ORFans -, appear in the genome of pandoraviruses was the main goal of our study.

As we investigated the most visible alterations of the otherwise nearly perfect collinearity of the p. celtis and p. quercus genomes, we identified 3 transposons of the hAT family (S_0, $S_{1,}$ and S_2 in **Figure 2**). The cargo of these mobile elements was found to be variable in gene number (10 for S_0, 11 for S_1, and 7 for S_2) and with a single overlap (pclt_cds_92, 681, 869) in addition to the transposases. These by-standing proteins exhibit no functional signature and have no homolog outside of the *Pandoraviridae* family. Interestingly, hAT family transposases have recently been identified in various Acanthamoeba species (Zhang et al., 2018). However, the gene contents of the p. celtis and p. quercus hAT transposons indicates that these mobile elements are not prominent vehicles of lateral gene transfers from the amoebal hosts to the pandoraviruses. With the exception of pclt_cds_870 (encoded by S_2), newly inserted transposons do not contribute genes unique to p. celtis. The hAT transposable elements only appear to ferry genes between pandoravirus strains, generating non-local duplications. Such exchanges might occur within an amoebal host undergoing multiple infections. To our knowledge, this is the first identification of hAT family transposons in DNA viruses. The fact that hAT transposons are not present in other well-documented families of large and giant virus infecting Acanthamoeba (Abergel et al., 2015; Aherfi et al., 2016; Fabre et al., 2017; Zhang et al., 2018) suggests that its transfer from host to virus is a rare event, or is specifically linked to pandoravirus infections.

This work identified 30 genes unique to p. celtis or p. quercus (**Table 1**) that we interpreted as encoding novel proteins that appeared after the recent divergence of these two strains from their common ancestor (i.e., below node 9 in **Figure 1**). These new genes are uniformly distributed along the p. celtis and p. quercus genomes, and do not co-localize with large genomic insertions or rearrangements, except for pqer_cds_665 and pqer_cds_673 encoded by the 30.5 kb segments unique to p. quercus (see Main Structural Features of the p. celtis and p. quercus Genomes). As noticed in our previous study (Legendre et al., 2018) all these proteins

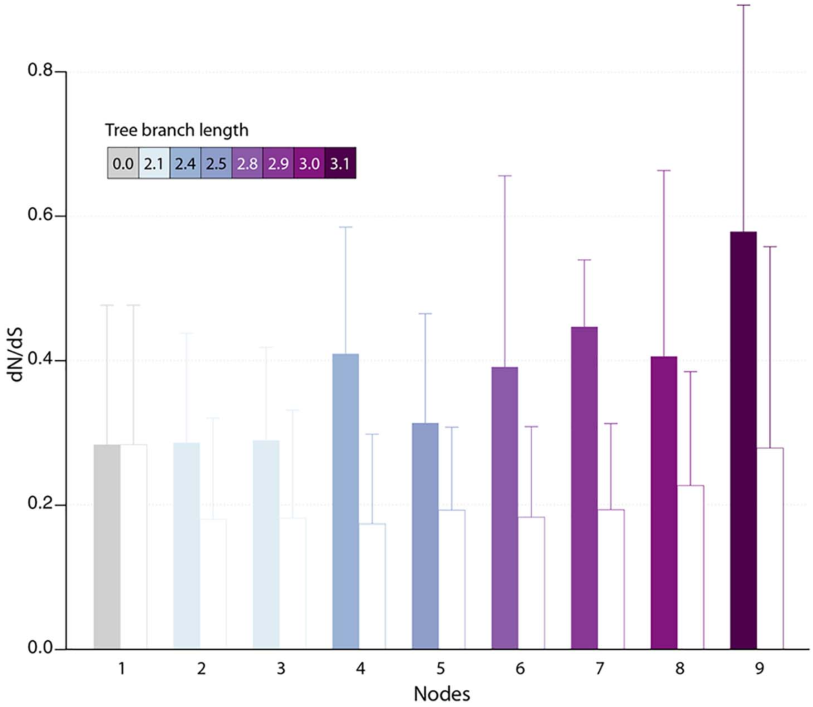


FIGURE 4 | Estimated selection pressure acting on pandoravirus genes as a function of their ancestry. Shown is the mean of dN/dS ratios (filled bars) for genes that are unique to the pandoravirus strains below a given node in their phylogeny tree (see **Figure 1**). Error bars correspond to standard deviations. Bars are colored according to the depth of the node in the tree. As a control, we calculated the dN/dS of pandoraviruses' core genes (empty bars). For a given node, we considered pairs of orthologous genes whose common ancestor correspond to that node.

are strict ORFans (not detected in any other organism, virus or other pandoravirus strains) and exhibit statistical features different from other pandoravirus genes, including a G+C content closer to that of non-coding/intergenic regions. We thus proposed that new proteins could be *de novo* created by the triggering of the transcription and translation of pre-existing non-coding sequences (the so-called "protogenes"). Here we further investigated this hypothesis by looking for eventual similarities of the new protein sequences with non-coding regions in other pandoraviruses at increasing evolutionary distances (**Figure 1**). For 8 of the 9 new p. celtis proteins and 15 of the 21 new p. quercus proteins we detected significant non-coding matches in various pandoravirus strains, in good agreement with their phylogenetic relationships. Trace of five new genes were only detected at the level of node 9, 8 at the more ancient node 8, 4 at the node 5, and 6 at the root (node 2) of the clade A pandoraviruses. Although these numbers are small and subject to fluctuations, such a distribution of hits suggests an evolutionary process combining a continuous spontaneous generation of open reading frames, followed by their drift back as non-coding sequences, unless they become transcriptionally active, and fixed as new proteins in a given strain. The maximum value of detected traces at node 8 might thus result from a compromise between the time interval required to generate *de novo* proteins, and the time during which they could retain a detectable similarity with the drifting non-coding regions from which they originated.

According to the above gene creation hypothesis, initially non-coding regions could act as precursors for new proteins, by the random opening of a suitable reading frame, followed by its transcriptional activation. Such a scenario is compatible with the distribution of new gene matches in other pandoravirus strains discussed above, where 8 correspond to intergenic regions, and 5 overlap with UTRs (**Table 1**). The flexibility of intergenic regions is further attested by the fact that 11 of the new genes have no matches. As shown in **Table 1**, a single new gene (pclt_cds_308) was found to co-localize with a ncRNA (pqer_ncRNA_47) despite the large number of ncRNAs detected in the p. celtis and p. quercus genomes. ncRNAs thus do not appear to be necessary intermediates in the emergence of new proteins. The low proportion of conserved ncRNAs between the two close strains (54%) suggests that the on/off expression status of non-coding genomic sequences is fluctuating very fast, and may even be variable at the viral population level.

The fact that a negative selection pressure (dN/dS < 1) is associated to the genes appeared since the divergence of p. celtis and p. quercus decisively reinforces the hypothesis of their *de novo* creation. Firstly, it provides the proof that these genes are truly expressed as *bona fide* proteins, given that differences between synonymous *vs.* non-synonymous mutations cannot be generated within non-translated nucleotide sequences. Any significant deviation from neutrality (dN/dS ≈ 1) can only be due to a selection process exerted on certain amino acids at certain positions of a true protein in order to improve, preserve, or eliminate its function.

The analysis of proteins whose creation appears most recent (node 9 in **Figure 4**) indicates a negative selection pressure whose (probably overestimated) average value (0.58) could correspond to functions slightly increasing the pandoravirus fitness. However, the interpretation of the selection pressure in the context of a function is a problem here, since we previously pointed out that it is highly unlikely that a newly created small protein will have a function (enzymatic, regulatory or structural) in the first place, even less a beneficial one (Legendre et al., 2018).

Thus, we propose an alternative interpretation, based on the knowledge accumulated over the years on the principles governing the stability of proteins. According to our scenario, most new proteins are from the translation of pre-existing non-coding "protogenes" (see Protein-Coding Genes Unique to p. celtis or p. quercus) and the rest from DNA inserts of unknown origin (such as pqer_cds_665 and pqer_cds_673). It is then unlikely that many of these new proteins lacking an evolutionary history will adopt a globular fold (Monsellier and Chiti, 2007; Watters et al., 2007; Tanaka et al., 2010). Most will be toxic by causing nonspecific aggregations within the host-virus proteome and be quickly eliminated through the reversal of these newborn viral genes to an untranscribed/untranslated state. In contrast, proteins whose folded 3-D structures do not prove to be toxic will enter an evolutionary process involving a negative selection pressure promoting the conservation of the amino acids responsible for their stability. Their proportion was estimated at about 34% for 100–200 residue proteins (Miao et al., 2004). In absence of an initial function, the other positions/residues will evolve in a neutral manner. Once “fixed,” the new genes will exhibit an average selection pressure lower than one, combining 1/3 of negative selection with 2/3 of neutrality. Following many studies that have concluded that efficient folding (Watters et al., 2007) and prevention of aggregation are important drivers of protein evolution (see Monsellier and Chiti, 2007), our scenario predicts that many genes specific to pandoraviruses should encode proteins not increasing their fitness, but whose stable 3-D structures may eventually serve as innovative platforms for new functions.

CONCLUSION

The detailed comparative analysis of p. celtis with its very close relative p. quercus enabled us to estimate the contributions of various microevolutionary processes to the steady increase of the pan-genome of the *Pandoraviridae* giant virus family. We first showed that large-scale genomic rearrangements (segmental duplications, translocations) are associated to transposable elements of the hAT family, widespread in metazoans, but until now unique to this family of viruses. However, these mobile elements mostly appear to shuffle pandoravirus genes between strains, without creating new ones or promoting host-to-virus horizontal gene transfers. In contrast with the popular view that horizontal gene transfer plays an important role in the evolution of large DNA viruses (Yutin and Koonin, 2013; Koonin et al., 2015; Schulz et al., 2017), this is definitely not a main cause of genome inflation in the *Pandoraviridae*, as we previously argued (Claverie and Abergel, 2013; Philippe et al., 2013; Abergel et al., 2015; Legendre et al., 2018). Finally, we also found that locally repeated regions are the siege of a competition between tandem duplication and gene deletion, locally reshaping the genomes without contributing to net genetic innovation (**Supplementary Figure S3**).

In continuity with our previous work on more distant pandoravirus strains, we found that the 30 protein-coding genes born since the divergence between the very close p. quercus and p. celtis strains were derived from preexisting non-coding sequences or small DNA segments of unknown origins inserted at randomly interspersed locations. We propose that random ORFs constantly emerge in non-coding regions and that their transcription is turned on in some pandoravirus strains, while they remain silent until they are deleted or diverge beyond recognition in others strains. Our results add strong support to the constant *de novo* creation of proteins, few of which are retained with a little initial impact on the virus fitness until eventually acquiring a selectable function.

Such a scenario, particularly visible and active in the *Pandoraviridae*, might also apply to the large proportion of ORFans encoded by other DNA viruses, from large eukaryotic viruses (Abergel et al., 2015) to much smaller bacteriophages, in which they can be the target of global functional studies (Berjón-Otero et al., 2017). Interestingly, the *de novo* gene creation scenario is gaining more and more acceptance beyond the realm of virology, recently to explain the origin of orphan protein even in mammals (Schmitz et al., 2018).

DATA AVAILABILITY

P. celtis’ annotated genome is deposited in GenBank under Accession n° MK174290.

AUTHOR CONTRIBUTIONS

CA, ML, J-MC, and YC designed the experiments. J-MA, AL, NP, SJ, and YC contributed to the data and performed the experiments. ML, SN, NT, OP, CA, and J-MC analyzed the data. ML, CA, and J-MC wrote the manuscript.

FUNDING

This work was partially supported by the French National Research Agency (ANR-14-CE14-0023-01), France Genomique (ANR-10-INSB-01-01), Institut Français de Bioinformatique (ANR–11–INSB–0013), the Fondation Bettencourt-Schueller (OTP51251), by a DGA-MRIS scholarship, and by the Provence-Alpes-Côte-d’Azur région (2010 12125). Proteomic experiments were partly supported by the Proteomics French Infrastructure (ANR-10-INBS-08-01) and Labex GRAL (ANR-10-LABX-49-01). We thank the support of the discovery platform and informatics group at EDyP and the PACA-Bioinfo platform.

ACKNOWLEDGMENTS

We kindly thank Hua-Hao Zhang for providing us with the hAT-transposon sequence he identified in the *Acanthamoeba castellanii* genome.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00430/full#supplementary-material

REFERENCES

Abergel, C., Legendre, M., and Claverie, J. M. (2015). The rapidly expanding universe of giant viruses: mimivirus, pandoravirus, pithovirus and mollivirus. *FEMS Microbiol. Rev.* 39, 779–796. doi: 10.1093/femsre/fuv037

Aherfi, S., Andreani, J., Baptiste, E., Oumessoum, A., Dornas, F. P., Andrade, A. C. D. S. P., et al. (2018). A large open pangenome and a small core genome for giant pandoraviruses. *Front. Microbiol.* 9:1486. doi: 10.3389/fmicb.2018.01486

Aherfi, S., Colson, P., La Scola, B., and Raoult, D. (2016). Giant viruses of amoebas: an update. *Front. Microbiol.* 7:349. doi: 10.3389/fmicb.2016.00349

Antwerpen, M. H., Georgi, E., Zoeller, L., Woelfel, R., Stoecker, K., and Scheid, P. (2015). Whole-genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba. *Genome Announc.* 3:e00136–15. doi: 10.1128/genomeA.00136-15

Atkinson, P. W. (2015). *hAT* transposable elements. *Microbiol. Spectr.* 3:MDNA3-0054-2014. doi: 10.1128/microbiolspec.MDNA3-0054-2014

Berjón-Otero, M., Lechuga, A., Mehla, J., Uetz, P., Salas, M., and Redrejo-Rodríguez, M. (2017). Bam35 tectivirus intraviral interaction map unveils new function and localization of phage ORFan proteins. *J. Virol.* 91:e0870–17. doi: 10.1128/JVI.00870-17

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Claverie, J. M., and Abergel, C. (2013). Open questions about giant viruses. *Adv. Virus Res.* 85, 25–56. doi: 10.1016/B978-0-12-408116-1.00002-1

Claverie, J. M., Abergel, C., and Legendre, M. (2018). Giant viruses that create their own genes. *Med. Sci.* 34, 1087–1091. doi: 10.1051/medsci/2018300

Corradi, N., Pombert, J. F., Farinelli, L., Didier, E. S., and Keeling, P. J. (2010). The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. *Nat. Commun.* 1:77. doi: 10.1038/ncomms1082

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16:157. doi: 10.1186/s13059-015-0721-2

Fabre, E., Jeudy, S., Santini, S., Legendre, M., Trauchessec, M., Couté, Y., et al. (2017). Noumeavirus replication relies on a transient remote control of the host nucleus. *Nat. Commun.* 8:15087. doi: 10.1038/ncomms15087

Floriano, A. M., Castelli, M., Krenek, S., Berendonk, T. U., Bazzocchi, C., Petroni, G., et al. (2018). The genome sequence of "*Candidatus Fokinia solitaria*": insights on reductive evolution in rickettsiales. *Genome Biol. Evol.* 10, 1120–1126. doi: 10.1093/gbe/evy072

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

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

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Koonin, E. V., Krupovic, M., and Yutin, N. (2015). Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Ann. N. Y. Acad. Sci.* 1341, 10–24. doi: 10.1111/nyas.12728

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* 23, 1026–1028. doi: 10.1093/bioinformatics/btm039

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

Latorre, A., and Manzano-Marín, A. (2017). Dissecting genome reduction and trait loss in insect endosymbionts. *Ann. N. Y. Acad. Sci.* 1389, 52–75. doi: 10.1111/nyas.13222

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.* 9:2285. doi: 10.1038/s41467-018-04698-4

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., et al. (2015). In-depth study of mollivirus sibericum, a new 30,000-y-old giant virus infecting Acanthamoeba. *Proc. Natl. Acad. Sci. U.S.A.* 112, E5327–E5335. doi: 10.1073/pnas.1510795112

Lopez-Madrigal, S., Latorre, A., Porcar, M., Moya, A., and Gil, R. (2011). Complete genome sequence of '*Candidatus Tremblaya princeps*' strain PCVAL, an intriguing translational machine below the living-cell status. *J. Bacteriol.* 193, 5587–5588. doi: 10.1128/JB.05749-11

Marchler-Bauer, A., and Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32, W327–W331. doi: 10.1093/nar/gkh454

McCutcheon, J. P., and Moran, N. A. (2011). Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* 10, 13–26. doi: 10.1038/nrmicro2670

Miao, J., Klein-Seetharaman, J., and Meirovitch, H. (2004). The optimal fraction of hydrophobic residues required to ensure protein collapse. *J. Mol. Biol.* 344, 797–811. doi: 10.1016/j.jmb.2004.09.061

Monsellier, E., and Chiti, F. (2007). Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO Rep.* 8, 737–742. doi: 10.1038/sj.embor.7401034

Philippe, N., Legendre, M., Doutre, G., Couté, 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. doi: 10.1126/science.1239181

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* 9, 173–175. doi: 10.1038/nmeth.1818

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Rocha, E. P. C., Smith, J. M., Hurst, L. D., Holden, M. T., Cooper, J. E., Smith, N. H., et al. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J. Theor. Biol.* 239, 226–235. doi: 10.1016/j.jtbi.2005.08.037

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., et al. (2019). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 47, D23–D28. doi: 10.1093/nar/gky1069

Schmitz, J. F., Ullrich, K. K., and Bornberg-Bauer, E. (2018). Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* 2, 1626–1632. doi: 10.1038/s41559-018-0639-7

Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., et al. (2017). Giant viruses with an expanded complement of translation system components. *Science* 356, 82–85. doi: 10.1126/science.aal4657

Szilágyi, A., and Skolnick, J. (2006). Efficient prediction of nucleic acid binding function from low-resolution protein structures. *J. Mol. Biol.* 358, 922–933. doi: 10.1016/j.jmb.2006.02.053

Tanaka, J., Doi, N., Takashima, H., and Yanagawa, H. (2010). Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. *Protein Sci.* 19, 786–795. doi: 10.1002/pro.358

Watters, A. L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., et al. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. *Cell* 128, 613–624. doi: 10.1016/j.cell.2006.12.04

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yutin, N., and Koonin, E. V. (2013). pandoraviruses are highly derived phycodnaviruses. *Biol. Direct.* 8:25. doi: 10.1186/1745-6150-8-25

Zhang, H. H., Zhou, Q. Z., Wang, P. L., Xiong, X. M., Luchetti, A., Raoult, D., et al. (2018). Unexpected invasion of miniature inverted-repeat transposable elements in viral genomes. *Mob. DNA* 9:19. doi: 10.1186/s13100-018-0125-4

Conflict of Interest Statement: 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.

frontiers in Microbiology

HYPOTHESIS AND THEORY
published: 19 March 2019
doi: 10.3389/fmicb.2019.00523

Viruses and Evolution – Viruses First? A Personal Perspective

Karin Moelling[1,2]* *and Felix Broecker*[3]

[1] *Institute of Medical Microbiology, University of Zurich, Zurich, Switzerland,* [2] *Max Planck Institute for Molecular Genetics, Berlin, Germany,* [3] *Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, United States*

The discovery of exoplanets within putative habitable zones revolutionized astrobiology in recent years. It stimulated interest in the question about the origin of life and its evolution. Here, we discuss what the roles of viruses might have been at the beginning of life and during evolution. Viruses are the most abundant biological entities on Earth. They are present everywhere, in our surrounding, the oceans, the soil and in every living being. Retroviruses contributed to about half of our genomic sequences and to the evolution of the mammalian placenta. Contemporary viruses reflect evolution ranging from the RNA world to the DNA-protein world. How far back can we trace their contribution? Earliest replicating and evolving entities are the ribozymes or viroids fulfilling several criteria of life. RNA can perform many aspects of life and influences our gene expression until today. The simplest structures with non-protein-coding information may represent models of life built on structural, not genetic information. Viruses today are obligatory parasites depending on host cells. Examples of how an independent lifestyle might have been lost include mitochondria, chloroplasts, *Rickettsia* and others, which used to be autonomous bacteria and became intracellular parasites or endosymbionts, thereby losing most of their genes. Even *in vitro* the loss of genes can be recapitulated all the way from coding to non-coding RNA. Furthermore, the giant viruses may indicate that there is no sharp border between living and non-living entities but an evolutionary continuum. Here, it is discussed how viruses can lose and gain genes, and that they are essential drivers of evolution. This discussion may stimulate the thinking about viruses as early possible forms of life. Apart from our view "viruses first", there are others such as "proteins first" and "metabolism first."

Keywords: evolution, RNA world, ribozymes, origin of life, viruses

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana-Champaign, United States

Reviewed by:
Guenther Witzany,
Telos-Philosophische Praxis, Austria
Herve Seligmann,
UMR MEPHI, Aix-Marseille Université,
France

****Correspondence:***
Karin Moelling
moelling@molgen.mpg.de

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

Received: *04 November 2018*
Accepted: *28 February 2019*
Published: *19 March 2019*

Citation:
Moelling K and Broecker F (2019)
Viruses and Evolution – Viruses First?
A Personal Perspective.
Front. Microbiol. 10:523.
doi: 10.3389/fmicb.2019.00523

RIBOZYMES AND VIROIDS

The origin of life on Earth has recently regained attention following the discovery of exoplanets within possible habitable zones (Kasting et al., 1993). The astronomical number of expected exoplanets suggests that there is a high statistical chance that life has also evolved somewhere else. This possibility stimulates the thinking of how life started on the early Earth, which may help to extrapolate to other planets.

Life presumably started simple. This is plausible, but is an assumption. The smallest known bacteria are still rather large. One of the smallest known metabolically autonomous bacterial species is *Pelagibacter ubique* with about 1,400 genes (Giovannoni et al., 2005). Genome reduction of

Mycoplasma mycoides by systematic deletion of individual genes resulted in a synthetic minimal genome of 473 genes (Hutchison et al., 2016). Can one consider simpler living entities?

There are elements with zero genes that fulfill many criteria for early life: ribozymes, catalytic RNAs closely related to viroids. They were recovered *in vitro* from 10^{15} molecules (aptamers), 220 nucleotides in length, by 10 rounds of selection. Among the many RNA species present in this collection of quasispecies RNAs were catalytically active members, enzymatically active ribozymes. The sequence space for 220-mer RNAs is about 3×10^{132} (Eigen, 1971; Wilson and Szostak, 1999; Brackett and Dieckmann, 2006).

The selected ribozymes were able to replicate, cleave, join, and form peptide bonds. They can polymerize progeny chemically, allow for mutations to occur and can evolve. One molecule serves as catalyst, the other one as substrate. Replication of ribozymes was demonstrated in the test tube (Lincoln and Joyce, 2009). Ribozymes can form peptide bonds between amino acids (Zhang and Cech, 1997). Thus, small peptides were available by ribozyme activity. Consequently, an RNA modification has been proposed as peptide nucleic acid (PNA), with more stable peptide bonds instead of phosphodiester bonds (Zhang and Cech, 1997; Joyce, 2002). Replication of RNA molecules can be performed chemically from RNA without polymerase enzymes. In addition, deoxyribozymes can form from ribonucleotides (Wilson and Szostak, 1999). Thus, DNA can arise from RNA chemically, without the key protein enzyme, the reverse transcriptase.

An entire living world is possible from non-coding RNA (ncRNA) before evolution of the genetic code and protein enzymes. Ribozymes naturally consist of circular single-stranded RNAs (Orgel, 2004). They lack the genetic triplet code and do not encode proteins. Instead, they exhibit structural information by hairpin-loops that form hydrogen bonds between incomplete double strands, and loops free to interact with other molecules. They represent a quasispecies in which many species of RNA may form, such as ribozymes, tRNA-like molecules, and other ncRNAs. RNAs within such a pool can bind amino acids. Ninety different amino acids have been identified on the Murchison meteorite found in Australia, while on Earth only about 20 of them are used for protein synthesis (Meierhenrich, 2008). Where formation of ribozymes occurred on the early Earth is a matter of speculation. The hydrothermal vents such as black smokers in the deep ocean are possibilities where life may have started (Martin et al., 2008). There, temperature gradients and clay containing minerals such as magnesium or manganese are available. Pores or niches offer possibilities for concentration of building blocks, which is required for chemical reactions to occur. Interestingly, also ice is a candidate for ribozyme formation and chemical reactions. Ice crystals displace the biomolecules into the liquid phase, which leads to concentration, creating a quasicellular compartmentalization where *de novo* synthesis of nucleotide precursors is promoted. There, RNA and ribozymes can emerge, which are capable of self-replication (Attwater et al., 2010).

tRNA-amino acid complexes can find RNAs as "mRNAs." Such interactions could have contributed to the evolution of the genetic code. This sequence of events can lead to primitive ribosome precursors. Ribozymes are the essential catalytic elements in ribosomes: "The ribosome is a ribozyme" (Cech, 2000), supplemented with about a hundred scaffold proteins later during evolution. The proteins have structural functions and contribute indirectly to enzymatic activity. Are these ribosome-bound ribozymes fossils from the early Earth? Small peptides can be formed by ribozymes before ribosomes evolved, whereby single or dimeric amino acids may originate from the universe (Meierhenrich, 2008).

Small peptides with basic amino acids can increase the catalytic activity of ribozymes as shown *in vitro* (Müller et al., 1994). Such proteins are known as RNA-binding proteins from RNA viruses that protect the RNA genome, with motifs such as RAPRKKG of the nucleocapsid NCp7 of HIV (Schmalzbauer et al., 1996). Peptides can enhance the catalytic activity of ribozymes up to a 100-fold (Müller et al., 1994). Such peptides of RNA viruses serve as chaperones that remove higher ordered RNA structures, allowing for more efficient interaction of RNA molecules and increasing transcription rates of RNA polymerases (Müller et al., 1994). Ribonucleoproteins may have also been functionally important during the evolution of ribosomes (Harish and Caetano-Anolles, 2012).

These pre-ribosomal structures are also similar to precursor-like structures of retroviruses. Reverse transcription can be performed by ribozymes chemically. This action does not necessarily require a protein polymerase such as the reverse transcriptase. Similarly, deoxyribonucleotides can arise by removal of an oxygen without the need of a protein enzyme (a reductase) as today, and allow for DNA polymerization (Wilson and Szostak, 1999; Joyce, 2002). The same elements of the precursors for ribosomes are also building blocks of retroviruses, which may have a similar evolutionary origin (Moelling, 2012, 2013). tRNAs serve as primers for the reverse transcriptase, and the sequence of promoters of transposable elements are derived from tRNAs (Lander et al., 2001). The ribozymes developed into more complex self-cleaving group II introns with insertion of genes encoding a reverse transcriptase and additional proteins (Moelling and Broecker, 2015; Moelling et al., 2017) (**Figure 1**).

It came as a surprise that the genomes of almost all species are rich in ncDNA, transcribed into ncRNAs but not encoding proteins, as evidenced, for instance, by the "Encyclopedia of DNA Elements" (ENCODE) project. ncDNA amounts to more than 98% of the human DNA genome (Deveson et al., 2017). Higher organisms tend to have more non-coding information, which allows for more complex modes of gene regulation. The ncRNAs are regulators of the protein-coding sequences. Highly complex organisms such as humans typically have a high number of ncRNA and regulatory mechanisms. ncRNA can range from close to zero in the smallest bacteria such as *Pelagibacter ubique* to about 98% in the human genome.

RNA viruses such as the retrovirus HIV harbor ncRNAs for gene regulation such as the *trans*-activating response element (TAR), the binding site for the Tat protein for early viral gene expression. Tat has a highly basic domain comprising mostly Lys and Arg residues, resembling other RNA binding proteins. ncRNA also serves on viral RNA genomes as ribosomal entry sites, primer binding sites or packaging signals. DNA synthesis depends on RNA synthesis as initial event, with RNA primers as starters for DNA replication, inside of cells as

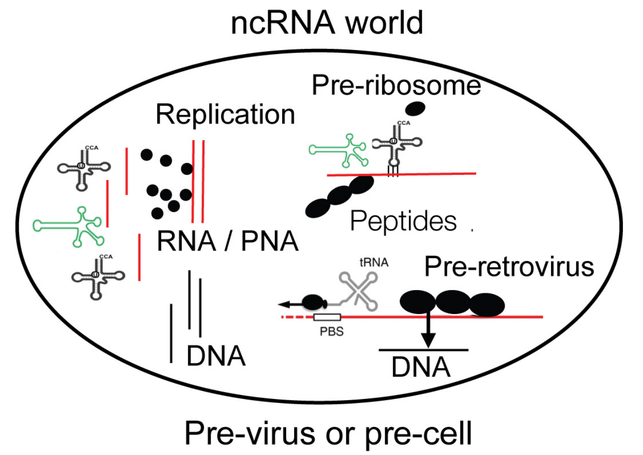


FIGURE 1 | A compartment is shown with essential components of life as discussed in the text. Non-coding RNA (ncRNA), ribozymes or viroids, can perform many steps for life without protein-coding genes but only by structural information. Individual amino acids are indicated as black dots and may be available on Earth from the universe. DNA may have existed before retroviruses. The compartment can be interpreted as pre-virus or pre-cell. Viroid, green; RNA, red; DNA, black.

well as during retroviral replication, proving a requirement of RNA (Flint, 2015).

The number of mammalian protein-coding genes is about 20,000. Surprisingly, this is only a fifth of the number of genes of bread wheat (Appels et al., 2018). Tulips, maize and other plants also have larger genomes, indicating that the number of genes does not necessarily reflect the complexity of an organism. What makes these plant genomes so large, is still an open question. Could the giant genomes possibly be the result to breeding of plants by farmers or gardeners?

According to Szostak there are molecules which appear like relics from the RNA world such as acetyl-CoA or vitamin B12, both of which are bound to a ribonucleotide for no obvious reason – was it "forgotten" to be removed? (Roberts and Szostak, 1997; Szostak et al., 2001; Szostak, 2011). Perhaps the connected RNA serves as structural stabilizer. Lipid vesicles could have formed the first compartments and enclosed ribozymes, tRNAs with selected amino acids, and RNA which became mRNA. Is this a pre-cell or pre-virus (**Figure 1**)?

Patel et al. (2015) demonstrated that the building blocks of life, ribonucleotides, lipids and amino acids, can be formed from C, H, O, P, N, S in a "one pot" synthesis. This study can be regarded as a follow-up study of the classical Urey-Miller *in vitro* synthesis of biomolecules (Miller, 1953; Miller and Urey, 1959). Transition from the RNA to the DNA world was promoted by the formation of the reverse transcriptase. The enzyme was first described in retroviruses but it is almost ubiquitous and found in numerous cellular species, many of which with unknown functions (Simon and Zimmerly, 2008; Lescot et al., 2016). It is an important link between the RNA and the DNA worlds. The name reverse transcriptase is historical and irritating because it is the "real" transcriptase during the transition from the RNA to the DNA world. Similarly, the ribonuclease H (RNase H) is an essential enzyme of retroviruses (Mölling et al., 1971). The RNase H turned out to be one of the five most frequent and ancient proteins (Ma et al., 2008) that belongs to a superfamily of more than sixty different unique representatives and 152 families with numerous functions (Majorek et al., 2014).

Some of the many tRNAs can become loaded with amino acids. There are viruses containing tRNA-like structures (TLS), resembling these early RNAs (Dreher, 2009). The TLS of these viruses typically bind to a single amino acid. TLS-viruses include plant viruses, such as Turnip yellow mosaic virus, in Peanut clump virus, Tobacco mosaic virus (TMV), and Brome mosaic virus. Only half a tRNA is found in Narnaviruses of fungi. The amino acids known to be components of tRNA-like viruses are valine, histidine and tyrosine. The structures were also designated as "mimicry," enhancing translation (Dreher, 2009, 2010). They look like "frozen" precursor-like elements for protein synthesis. This combination of a partial tRNA linked to one amino acid can be interpreted as an evolutionary early step toward protein synthesis, trapped in a viral element.

Ribozymes are related to the protein-free viroids. Viroids are virus-like elements that belong to the virosphere, the world of viruses (Chela-Flores, 1994). Viroids lack protein coats and therefore were initially not designated as viruses but virus-like viroids when they were discovered in 1971 by Theodor Diener. He described viroids as "living fossils" (Diener, 2016) (**Figure 2**).

From infected potatoes, Diener isolated the Potato spindle tuber viroid (PSTVd) whose genome was about a 100-fold smaller than those of viruses known at that time. The viroids known today are ranging from 246 to 467 nucleotides. They contain circular single-stranded RNA, are protein-free and self-replicating with no genetic information, but only structural

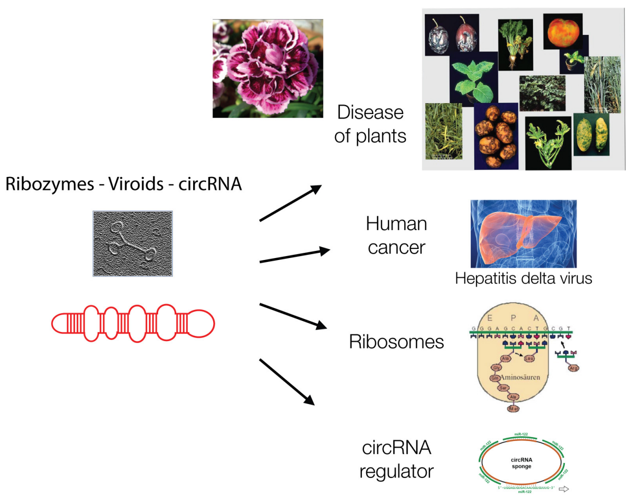


FIGURE 2 | Viroids are hairpin-loop structures and are shown schematically and as electron micrograph. Viroids are, like ribozymes, without genetic information and play major biological roles today in plant diseases, in carnation flowers, in liver cancer, as catalyst of protein synthesis in ribosomes and as circular regulatory RNAs, as "sponges" for other regulatory RNAs.

information in the form of hairpin-loops (Riesner et al., 1979). They can generate copies of themselves in the appropriate environment. They were designated as the "frontiers of life" (Flores et al., 2014).

The knowledge of virus composition was based on TMV and its crystallization by Wendell Stanley in 1935 (Pennazio and Roggero, 2000). The genome of TMV is protein-coding single-stranded RNA of about 6,400 nucleotides that is enclosed by a rod-like protein coat. Viroids, in contrast, do not encode proteins and lack coats but they are closely related to viruses. Viroids can lose their autonomy and rely on host RNA polymerases to replicate, are capable of infecting plants and many are economically important pathogens. There are two families, the nucleus-replicating *Pospiviroidae* such as PSTVd and the chloroplast-replicating *Avsunviroidae* like the Avocado sunblotch viroid (ASBVd). Their replication requires host enzymes. Thus, autonomy is replaced by dependence on host enzymes and an intracellular lifestyle.

Most viroids are often enzymatically active ribozymes – yet they are examples that this trait can get lost as a result of changing environmental conditions. Loss of ribozyme activity is a functional, not a genetic loss. Only the nuclear variants, the *Pospiviroidae*, can lose their ribozyme activity and use the cellular RNase III enzyme for their replication. In contrast, the *Avsunviroidae* are still active hammerhead ribozymes. Thus, inside the nucleus of a host cell, the enzymatic RNA function can become unnecessary. Not genes, but a function, the catalytic activity, gets lost.

Viroids did apparently not gain genes but cooperated for a more complex lifestyle. For example, Carnation small viroid-like RNA (CarSV RNA) cooperates with a retrovirus and is accompanied by a homologous DNA generated by a reverse transcriptase. This enzyme presumably originates from a pararetrovirus of plants. Pararetroviruses package virus particles at a different stage during replication than retroviruses, the DNA, not the RNA. This unique combination between two viral elements has so far only been detected with CarSV in carnation flowers (Flores et al., 2005, 2014). Why did such a cooperation evolve – perhaps by breeding gardeners? RNA is sensitive to degradation; therefore, genetic increase and growth of the genome may not be favorable energetically – at least not in plants. Gain of function is, in this case, cooperation.

The circular RNA (circRNA) is related to ribozymes/viroids as a chief regulator of other regulatory RNAs, a "sponge" absorbing small RNAs. Micro RNAs (miRNAs) are post-transcriptional regulators that are affected by the presence of circRNAs. circRNAs were detected in human and mouse brains and testes as well as in plants. They can bind 70 conserved miRNAs in a cell and amount up to 25,000 molecules (Hansen et al., 2013). Their structure is reminiscent of catalytically active ribozymes.

There is an exceptional viroid that gained coding information and entered the human liver (Taylor, 2009). The viroid is known as hepatitis delta virus (HDV). It has the smallest genome of any known animal virus of about 1,680 nucleotides. It has properties typical of viroids, since it contains circRNA, forms similar hairpin-loops and replicates in the nucleus using host enzymes. Two polymerases have to redirect their specificity from DNA to RNA to generate the HDV genome and antigenome. Both of them have ribozyme activity. In contrast to other ribozymes, HDV encodes a protein, the hepatitis delta antigen (HDVAg) that occurs in two forms, the small-HDVAg (24 kDa) supporting replication and the large-HDVAg (27 kDa) that helps virion assembly. The gene was presumably picked up from the host cell by recombination of HDV's mRNA intermediate with a host mRNA. Transmission depends on a helper virus, the Hepatitis B virus (HBV), which delivers the coat (Taylor, 2009) Does packaging by a helper virus protect the genome and thereby allow for a larger viroid to exist?

In plants, viroids may not be able to become bigger possibly due to their sensitivity to degradation – but they cannot become much smaller either. Only a single viroid is known that is completely composed of protein-coding RNA with triplets (AbouHaidar et al., 2014). Viroids and related replicating RNAs are error-prone replicating units and the error frequency imposes a certain minimal size onto them, as they would otherwise become extinct. This mechanism has been described as "error catastrophe," which prevents survival (Eigen, 1971, 2013). The viroids and related RNAs are the smallest known replicons. Smaller ones would become extinct in the absence of repair systems.

In summary, RNA can catalyze many reactions. Protein enzymes which may have evolved later have higher catalytic activities. Ribozymes are carriers of information, but do not require coding genes. Information is stored in their sequence and structure. Thus, replication of an initial RNA is followed by flow of information, from DNA to RNA to protein, as described the Central Dogma (Crick, 1968). Even an information flow from protein to DNA has been described for some archaeal proteins (Béguin et al., 2015). The DNA-protein world contains numerous ncRNAs with key functions. ncRNA may serve as a model compound for the origin of life on other planets. Hereby not the chemical composition of this molecule is of prime relevance, but its simplicity and multifunctionality. Furthermore, RNA is software and hardware in a single molecule, which makes it unique in our world. There are other scenarios besides the here discussed "virus-first," such as "protein-first", "metabolism-fist" or the "lipid world" (Segré et al., 2001; Andras and Andras, 2005; Vasas et al., 2010; Moelling, 2012). Some of these alternative concepts were built on phylogenomics, the reconstruction of the tree of life by genome sequencing (Delsuc et al., 2005). Surprisingly, it was Sir Francis Crick, one of the discoverers of the DNA double-helix, who stated that he would not be surprised about a world completely built of RNA. A similar prediction was made by Walter Gilbert (Crick, 1968; Gilbert, 1986). What a vision! Our world was almost 50 years later defined as "RNA-protein" world (Altman, 2013). One can speculate our world was built of ribozymes or viroids, which means "viruses first."

SPIEGELMAN'S MONSTER

ncRNAs appear as relics from the past RNA world, before DNA, the genetic code and proteins evolved. However, ncRNA is essential in our biological DNA world today. It is possible to produce such ncRNA today in the test tube by loss of genic information from protein-coding RNA. This reduction to ncRNA was demonstrated *in vitro* with phage RNA. Phage Qβ genomic RNA, 4,217 nucleotides in length, was incubated in the presence of Qβ replicase, free nucleotides and salts, a rich milieu in the test tube. The RNA was allowed to replicate by means of the Qβ replicase. Serial transfer of aliquots to fresh medium led to ever faster replication rates and reduction of genomic size, down to 218 nucleotides of ncRNA in 74 generations. This study demonstrated that, depending on environmental conditions, an extreme gene reduction can take place. This experiment performed in 1965 was designated as "Spiegelman's Monster." Coding RNA became replicating ncRNA (Spiegelman et al., 1965; Kacian et al., 1972)!

Manfred Eigen extended this experiment and demonstrated further that a mixture containing no RNA to start with but only ribonucleotides and the Qβ replicase can under the right conditions in a test tube spontaneously generate self-replicating ncRNA. This evolved into a form similar to Spiegelman's Monster. The presence of the replicase enzyme was still necessary in these studies. Furthermore, a change in enzyme concentration and addition of short RNAs or an RNA intercalator influenced the arising RNA population (Sumper and Luce, 1975; Eigen, 2013). Thus, the complexity of genomes depends on the environment: poor conditions lead to increased complexity and rich environments to reduced complexity.

The process demonstrated in this experiment with viral components indicates that reversion to simplicity, reduction in size, loss of genetic information and speed in replication can be major forces of life, even though this appears to be like a reversion of evolution. The experiment can perhaps be generalized from the test tube to a principle, that the most successful survivors on our planet are the viruses and microorganisms, which became the most abundant entities. Perhaps life can start from there again.

These studies raise the question of how RNA molecules can become longer, if the small polymers become smaller and smaller, replicate faster and outcompete longer ones. This may be overcome by heat flow across an open pore in submerged rocks, which concentrates replicating oligonucleotides from a constant feeding flow and selection for longer strands. This has been described for an increase from 100 to 1,000 nucleotides *in vitro*. RNA molecules shorter than 75 nucleotides will die out (Kreysing et al., 2015). Could a poor environment lead to an increase of complexity? This could be tested. Ribozymes were shown to grow in size by uptake of genes, as demonstrated for HDV (Taylor, 2009).

MICROBIOME IN THE HUMAN INTESTINE

An interesting recent unexpected example supporting the notion that environmental conditions influence genetic complexity, is the human gut microbiome. Its complexity increases with diverse food, while uniform rich food reduces its diversity and may lead to diseases such as obesity. Colonization of the human intestinal tract starts at birth. A few dozen bacterial and viral/phage species are conserved between individuals (core sequences) as a stable composition (Broecker et al., 2016c, 2017). Dysbiosis has been observed in several chronic diseases and in obesity, a loss of bacterial richness and diversity. Nutrition under affluent conditions with sugar-rich food contributes to obesity, which results in a significant reduction of the complexity of the microbiome. This reduction is difficult to revert (Cotillard et al., 2013; Le Chatelier et al., 2013). The gut microbiome in human patients with obesity is reminiscent of the gene reduction described in the Spiegelman's Monster experiment: reduction of genes in a rich environment.

The reduction of the complexity of the microbiome is in part attributed to the action of phages, which under such conditions, defined as stress, lyse the bacteria. Fecal microbiota transplantation can even be replaced by soluble fractions containing phages or metabolites from the donor without bacteria (Ott et al., 2017). Analogously, the most highly complex microbiomes are found in indigenous human tribes in Africa, which live on a broad variety of different nutrients. It is a slow process, though, to increase gut microbiota complexity by diverse nutrition. The obesity-associated microbiota that survive are fitter and more difficult to counteract. Urbanization and westernization of the diet is associated with a loss of microbial biodiversity, loss of microbial organisms and genes (Segata, 2015).

To understand the mechanism and driving force for genome reduction, deletion rates were tested by insertion of an indicator gene into the *Salmonella enterica* genome. The loss of the indicator gene was monitored by serial passage in rich medium. After 1,000 generations about 25% of the deletions caused increased bacterial fitness. Deletions resulted in smaller genomes with reduced or absence of DNA repair genes (Koskiniemi et al., 2012). Gene loss conferred a higher fitness to the bacteria under these experimental conditions.

MIMIVIRUSES – GAIN OR LOSS?

The recently discovered mimiviruses and other giant viruses are worth considering for understanding the evolution of life with respect to the contribution of viruses. Their hosts are, for example, *Acanthamoeba*, *Chlorella,* and *Coccolithus* algae (*Emiliania huxleyi*), but also corals or sponges as discussed more recently. Mimiviruses were first discovered in cooling water towers in Bradford, United Kingdom in 2003 with about 1,000 genes, most of which unrelated to previously known genes. Mimiviruses have received attention because they contain elements that were considered hallmarks of living cells, not of viruses, such as elements required for protein synthesis, tRNAs and amino acid transferases. The mimiviruses harbor these building blocks as incomplete sets not sufficient for independent protein synthesis as bacteria or archaea can perform, preventing them from leading an autonomous life (La Scola et al., 2003, 2008). They are larger than some bacteria. Giant viruses can be looked at as being on an evolutionary path toward a cellular organism. Alternatively, they may have evolved from a cellular organism by loss of genetic information (Nasir and Caetano-Anolles, 2015). Giant viruses have frequently taken up genes from their hosts by horizontal gene transfer (HGT) (La Scola et al., 2008; Nasir and Caetano-Anolles, 2015; Colson et al., 2018). A graph on genome sizes shows that mimiviruses and bacteria overlap in size, indicating a continuous transition between viruses and bacteria and between living and non-living worlds (based on Holmes, 2011) (**Figure 3**). Other giant viruses, such as megaviruses, were discovered in the ocean of Chile with 1,120 genes. Most recently the Klosneuvirus was identified in the sewage of the monastery Klosterneuburg in Austria in 2017 with 1.57 million (mio) basepairs (Mitch, 2017). *Pithovirus sibericum* is the largest among giant viruses discovered to date with a diameter of 1.5 microns, a genome of 470,000 bp with 467 putative genes, 1.6 microns in length, and it is presumably 30,000 years old as it was recovered from permafrost in Siberia (Legendre et al., 2014). The smaller *Pandoraviruses* with 1 micron in length have five times larger genomes, 2,500,000 bp (Philippe et al., 2013) (**Figure 3**).

The giant viruses can even be hosts to smaller viruses, the virophages, reminiscent of bacteriophages, the viruses of bacteria. These virophages such as Sputnik are only 50 nm in size with 18,343 bp of circular dsDNA and 21 predicted protein-coding genes. They replicate in viral factories and consume the resources of the mimivirus, thereby destroying it. Some, virophages can even integrate into the genome of the cellular host and can be reactivated when the host is infected by giant viruses. Thus, giant viruses suggest that viruses are close to living entities or may have been alive (La Scola et al., 2008; Fischer and Hackl, 2016).

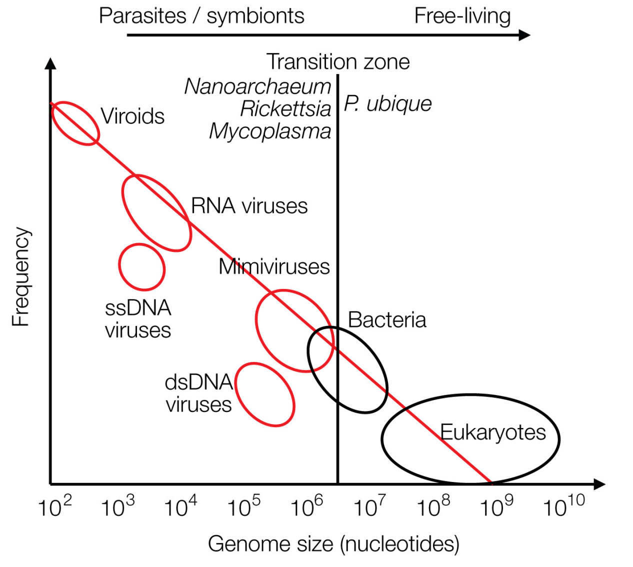


FIGURE 3 | Size distribution of viruses (red circles), free-living bacteria and eukaryotes (black circles) are shown relative to their frequencies. The transition zone between parasites or symbionts versus free-living species is indicated by a black line. The transition is not a sharp borderline as shown by the circles and as discussed in the text (modified from Holmes, 2011).

In biology it is common to distinguish between living and dead matter by the ability to synthesize proteins and replicate autonomously. The giant viruses may be considered as missing link between the two, because they harbor "almost" the protein synthesis apparatus. The transition from living to the non-living world is continuous, not separated by a sharp borderline (**Figure 3**).

Viruses are not considered alive by most of the scientific community and as written in textbooks, because they cannot replicate autonomously. Yet some of the giant viruses are equipped with almost all components of the protein synthesis machinery close to bacteria suggesting that they belong to the living matter (Schulz et al., 2017). The ribozymes may have been the earliest replicating entity. Perhaps also other viruses were initially more independent of the early Earth than they are today. As described in **Figure 1** there may have been initially no major difference between an early virus or an early cell. Only later viruses may have given up their autonomous replication and became parasites – as has been described for some bacteria (see below).

Efforts have been made to identify the smallest living cell that is still autonomously replicating. Among the presumably smallest naturally occurring bacteria is *Pelagibacter ubique* of the SAR11 clade of bacteria (Giovannoni, 2017), which was discovered in 1990. It is an alpha-proteobacterium with 1,389 genes present ubiquitously in all oceans. It can reach up to 10^{28} free living cells in total and represents about 25% of microbial plankton cells. Very little of its DNA is non-coding. It harbors podophage-type phages, designated as "pelagiphage" (Zhao et al., 2013). This small bacterium was designated as the most common organism on the planet. Why is it so successful? This autonomous bacterium is smaller than some parasitic giant viruses. Craig Venter, who first succeeded in sequencing the human genome, tried to minimize the putative smallest genome of a living species, from *Mycoplasma mycoides*, a parasitic bacterium that lives in ruminants (Gibson et al., 2008, 2010). His group synthesized a genome of 531,000 bp with 473 genes, 149 of them (32%) with unknown functions (Hutchison et al., 2016). Among the smallest parasitic living organisms is *Nanoarchaeum equitans*. It is a thermophile archaeon which lives at 80°C and at pH 6 with 2% salt (Huber et al., 2003). Its genome has a size of 490,000 bp and encodes 540 genes. *N. equitans* is an obligate symbiont of a bigger archaeon, *Ignicoccus* riding on it as on a horse, hence the name (Huber et al., 2003).

The world of viruses covers a range of three logs in size of their genomes: from zero genes to about 2,500 genes amounting to about 2,500,000 bp of DNA. The zero-gene viroids are about 300 bases in length (**Figure 3**).

The virosphere is the most successful reservoir of biological entities on our planet in terms of numbers of particles, speed of replication, growth rates, and sequence space. There are about 10^{33} viruses on our planet and they are present in every single existing species (Suttle, 2005).

There is no living species without viruses! Viruses also occur freely in the oceans, in the soil, in clouds up to the stratosphere and higher, to at least 300 km in altitude. They populate the human intestine, birth canal, and the outside of the body as protective layer against microbial populations. Microbes contain phages that are activated during stress conditions such as lack of nutrients, change in temperatures, lack of space and other changes of environmental conditions.

RETROVIRUSES AS DRIVERS OF EVOLUTION

One of the most earth-shaking papers of this century was the publication of the human genome sequence (Lander et al., 2001). About half, possibly even two-thirds of the sequence are composed of more or less complete endogenous retroviruses (ERVs) and related retroelements (REs) (de Koning et al., 2011). REs amplify via copy-and-paste mechanisms involving a reverse transcriptase step from an RNA intermediate into DNA. In addition, DNA transposable elements (TEs) move by a cut-and-paste mechanism. The origin of REs is being discussed as remnants of ancient retroviral germline infections that became evolutionarily fixed in the genome. About 450,000 human ERV (HERV) elements constitute about 8% of the human genome consisting of hallmark retroviral elements like the *gag, pol, env* genes and flanking long terminal repeats (LTR) that act as promoters (Lander et al., 2001). Howard Temin, one of the discoverers of the reverse transcriptase, in 1985 already described endogenous retrovirus-like elements, which he estimated to about 10% of the human and mouse genome sequence (Temin, 1985). The actual number is about 45% as estimated today (Lander et al., 2001). In some genes such as the *Protein Kinase Inhibitor B* (*PKIB*) gene we determined about 70% retrovirus-related sequences (Moelling and Broecker, 2015). Is there a limit? Could it have been 100%? Retroviruses are estimated to have entered the lineage of the mammalian genome 550 million years ago (MYA) (Hayward, 2017). Older ERV sequences may exist but are unrecognizable today due to the accumulation of mutations.

ERVs undergo mutations, deletions or homologous recombination events with large deletions and can become as short as solo LTR elements, which are a few hundred bp in length – the left-overs from full-length retroviral genomes of about 10,000 bp. The LTR promoters can deregulate neighboring genes. Homologous recombination events may be considered as gene loss or gene reduction events. It is the assumption that the ERVs, which were no longer needed for host cell defense, were no longer selected for by evolution and consequently deleted as unnecessary consumers of energy.

Eugene Koonin points out that infection and integration are unique events occurring at a fast pace, while loss and gene reduction may take much longer time frames (Wolf and Koonin, 2013).

A frequent gene reduction of eukaryotic genomes is the loss of the viral envelope protein encoded by the *env* gene. Without a coat, retroviruses can no longer leave the cell and infect other cells. They lose mobility and become obligatory intracellular elements. Helper viruses can supply envelope proteins in *trans* and mobilize the viruses. TEs or REs can be regarded as examples of coat-free intracellular virus relics – or could it have been the other way round, perhaps precursors of full-length retroviruses? These elements can be amplified intracellularly and modify the host genomes by integration with the potential danger of gene disruption and genetic changes. REs can lead to gene duplications and pseudogene development, with one copy for stable conservation of acquired functions and the other one for innovations (Cotton and Page, 2005). Such duplications constitute large amounts of mammalian genomes (Zhang, 2003). Retroviruses have an RNase H moiety duplication, one of which serves as a catalytically inactive linker between the RT polymerase and the enzymatically active RNase H (Xiong and Eickbush, 1990; Malik and Eickbush, 2001; Moelling and Broecker, 2015; Moelling et al., 2017). This gene duplication dates back to 500 mio years (Cotton and Page, 2005).

Gene duplications are a common cause of cancer, which often occurs only in the genome of the cancer cell itself, less affecting offsprings. Myc, Myb, ErbB2, Ras, and Raf are oncogenes amplified in diverse types of human cancers (Vogelstein and Kinzler, 2002). The ability of retroviruses to integrate makes them distinct from endosymbionts which stay separate. Yet the net result is very similar, acquisition of new genetic information, which is transmitted to the next generation, if the germline is infected and endogenization of the virus occurred.

Viral integration is not limited to eukaryotic cells but also a mechanism in prokaryotes for maintenance of the lysogenic state of phages inside bacteria.

Also, for other eukaryotic viruses such as HBV, the envelope surface antigen BHsAg can be deleted, which leads to an obligatory intracellular life style for the virus, which especially in the presence of HCV promotes cancer (Yang et al., 2016).

HIV has been shown to rapidly lose one of its auxiliary genes, *nef*, originally for negative factor. The gene was lost within a rather low number of passages of the virus grown under tissue culture conditions by selection for high virus titer producing cells. Deletion of *nef* resulted in a significant increase of the virus titer in culture – hence the name. The *nef* gene product was of no need inside tissue culture cells, rather it was inhibitory for replication. However, it is essential for pathogenicity in animals, and subsequently *nef* was reinterpreted as "necessary factor" (Flint, 2015).

Also, the human hosts of HIV can lose a significant terminal portion of a seven transmembrane receptor in lymphocytes, the primary target cell for HIV entry and for virus uptake. This molecule, the CCR5 cytokine receptor is truncated by 32 carboxy-terminal amino acids (CCR5-Δ32), disabling the receptor functionally. The allele frequency of the mutant CCR5-Δ32 mutant is about 10% in the European population, making these people resistant to HIV infections (Solloch et al., 2017). This gene loss in Europeans has been shown to make the individuals resistant not only against HIV infection but also against malaria. This may have been the selective pressure in the past before HIV/AIDS arose. No side effect for humans lacking this gene has been described (Galvani and Slatkin, 2003).

Viruses have been proven to be drivers of evolution (Villarreal and Witzany, 2010), including the human genome, which by at least 45% is composed of sequences related to retroviruses. In addition, endogenized retroviruses supplied the syncytin genes that are essential for the development of the mammalian placenta, and allowed the growth of embryos without its rejection by the maternal immune system (Dupressoir et al., 2012). Thus, the same property which causes immunodeficiency in HIV-infected patients and leads to AIDS causes syncytia formation, cell fusion after infection by a retrovirus. Viruses have also been proposed to be at the origin of the evolution of adaptive immunity (Villarreal, 2009). Thus, viruses shaped genomes by supplying essential genes and mechanisms.

ENDOGENIZATION OF VIRUSES

Endogenization of retroviruses has occurred in the mammalian genomes for at least 550 mio years (Hayward, 2017). If the integrated ERVs did not provide any selective advantage, they deteriorated and accumulated mutations with loss of function. This was directly proven by reconstruction of an infectious retrovirus from the consensus sequence of 9 defective endogenous virus sequences, designated as Phoenix. The virus was expressed from a constructed synthetic DNA clone in cell culture and formed virus particles identified by high resolution microscopic analysis (Dewannieux and Heidmann, 2013).

The koalas in Australia are currently undergoing endogenization of a retrovirus (koala retrovirus, KoRV) in "real time" and demonstrate possible consequences for immunity. In the early 1900s, some individuals were transferred to islands, including Kangaroo Island, close to the Australian mainland for repopulation purposes, as koalas were threatened to become extinct. Today, the majority of the koala population is infected by KoRV, which is closely related to the Gibbon ape leukemia virus (GALV). Yet, koalas isolated on Kangaroo Island are KoRV negative, which allows dating the introduction of KoRV into the koala population to about one hundred years ago. Many of the infected koalas fell ill and died, yet some populations became resistant within about 100 years, corresponding to about 10 generations. The koalas likely developed resistance due to the integrated DNA proviruses. The retrovirus is transmitted as exogenous as well as endogenous virus, similar to the Jaagsiekte sheep retrovirus (JSRV), whereby the endogenized viruses protect with a viral gene product, such as Env, against *de novo* infections by "superinfection exclusion" (Tarlinton, 2012). The contribution of retroviruses to the antiviral defense is striking, since all retroviral genes have analogous genes in the siRNA/RNAi defense mechanism of eukaryotic cells (Moelling et al., 2006).

VIRUSES PROTECT AGAINST VIRUSES

Retroviruses can protect against infection by other related viruses, for example, by expressing Env proteins that block cell-surface receptors (Villarreal, 2011). A comparable mechanism protects bacterial cells against DNA phages, by integrated phage DNA fragments that are transcribed into mRNA and hybridize to incoming new DNA phages and thereby lead to their destruction by hybrid-specific nucleases, CRISPR/Cas immunity (Charpentier and Doudna, 2013). It is often not realized that immunity acquisition in bacteria and mammalian cells follow analogous mechanisms (**Figure 4**).

Integration of retroviruses normally occurs in somatic cells after infection as an obligatory step during the viral life cycle. Infection of germline cells can lead to transmission to the next generation and ultimately result in inherited resistance. Endogenized retroviruses likely caused resistance

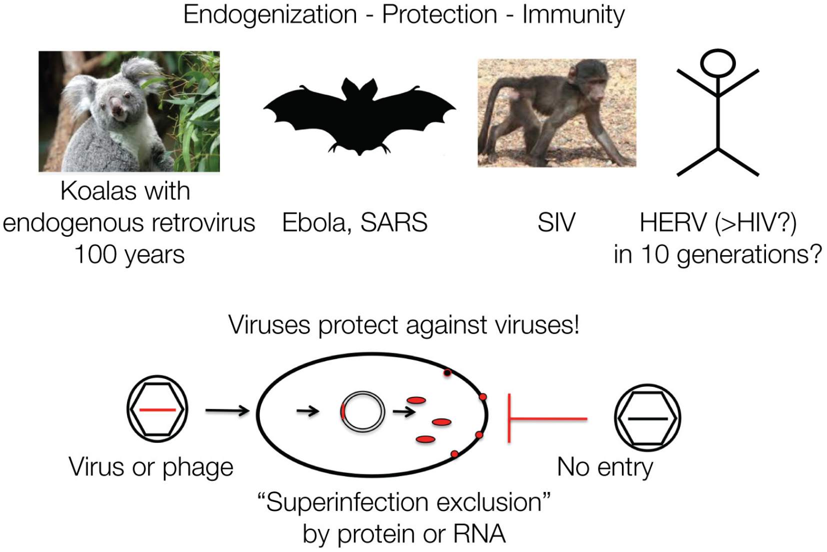

Endogenization - Protection - Immunity
Koalas with endogenous retrovirus 100 years
Ebola, SARS
SIV
HERV (>HIV?) in 10 generations?
Viruses protect against viruses!
Virus or phage
No entry
"Superinfection exclusion" by protein or RNA

FIGURE 4 | Viruses protect against viruses: retroviruses protect a cell against a new infection by a similar virus designated as "superinfection exclusion" or viral interference. This is mediated by viral gene products such as proteins or nucleic acids. Similarly, phages protect against phages: superinfection of bacteria is prevented by CRISPR/Cas RNA originating from previous infections. The mechanisms of defense against viruses and phages are analogous. Protection by viruses or phages against superinfections represents cellular defense and acquired immunity. The four examples are discussed in the text.

to the exogenous counterparts. Similarly, resistance to Simian Immune Deficiency virus (SIV) in some monkey species may be explained by endogenization (Li et al., 2017, 2018). In the case of phages and their prokaryotic hosts the mechanism is described as CRISPR/Cas, which follow analogous principles of "endogenization" of incoming genetic material for subsequent exclusion.

One may speculate that HIV may also eventually become endogenized into the human genome. There is some evidence that HIV can infect human germline cells and can be transmitted to the embryonic genome (Wang et al., 2011). How long this may take is not known – 10 generations?

The loss of function of ERVs can occur by mutations, deletions of the *env* or other genes and ultimately all coding genes by homologous recombination, leaving behind only one LTR. The number of retrovirus-like elements add up to about 450,000, corresponding to 8% of the human genome (Lander et al., 2001; Cordaux and Batzer, 2009). The promoter regions were analyzed for their contribution to cancer by activating neighboring genes – as a consequence of a former retrovirus infection. Indeed, activated cellular genes by "downstream promotion" were identified in animal studies with activation of the *myc* gene as one of many examples, leading to chronic, not acute development of cancer (Ott et al., 2013). As a general mechanism for human cancer today the LTRs are, however, not identified as a major culprit. Most of the ERVs we find today have been integrated during evolution in introns or other regions where their presence is relatively harmless. Did the other ones result in death of the carriers which disappeared? The effects of LTRs on the expression levels of neighboring host genes was studied with the endogenous human virus, HERV-K, as a possible cause of cancer, but this appears not to be a general phenomenon (Broecker et al., 2016b). As shown for the koalas, ERVs can confer immunity to viral infections (Feschotte and Gilbert, 2012).

A related ERV, HERV-H, was shown to produce an RNA that keeps early embryonic cells pluripotent and even revert adult cells to regain pluripotency (Grow et al., 2015). Thus, the role of ERVs may be more complex than we presently know.

Transposable elements and REs that lost the ability of cellular transmission by deletion of the coat protein majorly contribute to genetic complexity of host cells. They are "locked" inside the cells and are major drivers of the increase of genetic complexity (Cordaux and Batzer, 2009). One could speculate that these intracellular elements are replication-incompetent retroviruses lacking coats (Lander et al., 2001). Bats transmit viruses such as Ebola and SARS coronavirus without suffering from disease (Beltz, 2018). Even RNA viruses such as Bornaviruses have been shown to integrate by illegitimate reverse transcription, possibly also supplying immunity against superinfection (Katzourakis and Gifford, 2010).

ENDOSYMBIOSIS

There are two prominent events that significantly contributed to the success of life and the formation of cells. Both of them are associated with gene reduction. This phenomenon may play a role for the evolution of viruses from autonomous to parasitic lifestyles. In the 1960s Lynn Margulis proposed an extracellular origin for mitochondria (Margulis, 1970, 1993). An ancestral cell, perhaps an archaeon, was infected by an anaerobic bacterium, which gave rise to mitochondria. Similarly, cyanobacteria formed the chloroplasts in modern plant cells. Mitochondria arose around 1.45 billion years ago (BYA) (Embley and Martin, 2006). Mitochondria and chloroplasts are the most striking examples for a change in lifestyle from autonomous bacteria to endosymbionts. This transition is often considered as extremely rare and a hallmark of evolution of life on our planet. However, there are many other obligate intracellular parasites such as *Rickettsia*, *Chlamydia trachomatis, Coxiella burnetii* (the causative agent of Q fever), *Mycobacterium leprae*, *M. tuberculosis*, and *M. mycoides* (Beare et al., 2006).

The change of lifestyle of the endosymbionts in the two cases of mitochondria and chloroplasts is striking. Both of them drastically reduced their genetic make-up. Mitochondria contain less than 37 genes, left from the original about 3,000 genes. Is endogenization of retroviruses, the ERVs, which are integrated into germline cells, related to endosymbiosis? Are these endosymbionts models for the transition from autonomous lifestyle to a parasitic life- which may have taken place with viruses?

A more recent typical example for a reductive evolution are *Rickettsia.* These bacteria were assumed for some time to be viruses because of their obligatory intracellular parasitic existence. *Rickettsia* have evolved from autonomously replicating bacteria. Reductive evolution of endosymbionts can yield bacteria with tiny genomes on the expense of autonomous extracellular life. Their genomes are 1.11 mio bp in length with about 834 protein-coding genes, and loss of 24% by reductive evolution (Ogata et al., 2001). *Rickettsia* may have some relationship with cyanobacteria, which are considered as the major symbionts.

Can one speculate that viruses may have been autonomous entities initially? Viroids may have undergone transition from autonomy to parasites, just as shown for mitochondria, chloroplasts or *Rickettsia*? To which extent have viruses been autonomous and independent of cellular metabolisms originally – and contributed to the origin of cells? Could they only later have lost their autonomy and become parasitic?

VIRUSES AND ONCOGENES

Viruses are minimalistic in their composition and must have undergone stringent gene reductions (Flint, 2015). How small can their genomes become? Most coding RNA viruses still contain regulatory elements, ncRNA at the 3′ and 5′ terminal regions for ribosomal entry, protein synthesis, transcriptional regulation, and others.

A subgroup of retroviruses is an interesting example in respect to simultaneous loss and gain of genetic information. The oncogenic retroviruses or tumorviruses can recombine with cellular genes which under the promoters of retroviruses can become oncogenes and drivers of cancer. About a hundred oncogenes have been selected for in the laboratories and studied over decades for understanding the molecular mechanisms of cancer. Selection for growth advantages of the host cells led to the discovery of the fastest growth-promoting oncogenes we know today, such as Ras, Raf, ErbB or Myc, which are in part successful targets for anticancer drugs (Moelling et al., 1984).

These oncogenes were in most cases taken up by the retroviruses at the expense of structural (*gag*), replicating (*pol*) or envelope (*env*) genes, and are often expressed as fusion proteins with Gag. Thus, oncogenic retroviruses are obligatory intracellular defective viruses and were selected for in the laboratory by researchers for the oncogenes with the most potent growth promoting ability. They need the supply of replicatory genes in *trans* from co-infecting helper viruses to infect other cells (Flint, 2015). Retroviruses are able to pick up cellular genes, transfer and integrate them into neighboring cells. Some strains of Rous sarcoma virus maintain replication competent when carrying the cell-derived *src* (for sarcoma) oncogene encoding a protein of 536 amino acids that apparently can fit into the retroviral particle along with the full-size viral genome (Broecker et al., 2016a). Spatial reasons may have influenced the formation of oncogenic retroviruses and limited their size and thereby led to their defective phenotypes.

There are indications that the uncontrolled activity of (retro)transposons in germline cells can result in diseases such as male infertility – presumably by "error catastrophe," caused by too many transposition events. In mammals, piRNAs tame transposon activity by means of the RNase H activity of PIWI proteins during spermatogenesis (Girard et al., 2006).

Only a minority of viruses are pathogens; most of them do not cause diseases. On the contrary, they are most important as drivers of evolution, as transmitters of genetic material, as innovative agents. In particular, the RNA viruses are the most innovative ones. Some of them are pathogenic and dangerous, such as HIV or influenza virus, or viroids in plants. RNA viruses are able to change so rapidly that the host immune system is unable to counteract the infection. Pathogenicity arises when environmental conditions change, for instance, when a virus enters a new organism or species.

Increase of cellular complexity by viruses is an important feature of evolution. Such major evolutionary changes are recently taken as arguments against the evolutionary theory by Charles Darwin who considered gradual changes, small increments by mutations as the main basis for selection and evolution. New criticism is addressing this thinking, considering larger changes as evolutionary drivers. Such changes arise by many complex phenomena such as endosymbiosis, infection by prokaryotes, viruses and fungi, recombination of genes, HGT, infections, sex. Dramatic changes such as endosymbiosis or pathogen infections extend Darwin's concept of evolution.

CONCLUSION

There are numerous examples for the contribution of viruses to the evolution of life since at least as long as 550 MYA (Hayward, 2017). But genetic noise through random mutations does not allow us to go back to the origin of life. It may not be impossible that the earliest compartment was indistinguishable, either a pre-cell or a pre-virus. By analogy one may speculate that at some point autonomous viruses gave up independence for an obligatory intracellular life – as has been described for mitochondria and chloroplasts but also intracellular bacteria such as *Rickettsia*. This speculation is based on the concept that early life must have started simple and with high genetic variability and then became more complex. But complexity can be given up for a less energy consuming lifestyle with small genomes and high speed of replication (Moelling, 2012, 2013). Therefore, the question may be repeated: "Are viruses our oldest ancestors?" Some fossil life can be partially reproduced *in vitro* by Spiegelman's Monster and Eigen's follow-up experiments, explaining the great surviving potential of simple ncRNA.

Viruses can be pathogens, but their recognition as primarily causing diseases is wrong. This notion is based on the history of viruses in medicine, as explained in a book entitled "Viruses: More Friends Than Foes" (Moelling, 2017). The scenario described here focuses on viruses as drivers of evolution.

The early RNA world gained interest 20–30 years ago as evidenced by the references provided above. Surprisingly, there are scientists who still believe in the "pansperm hypothesis" and think that retroviruses are of extraterrestric origin (Steele et al., 2018). The recent interest in the origin of life arose from the newly discovered exoplanets whose number increases daily – and which may be as numerous as 10^{25}. Thus, pure statistics make some people believe that there is extraterrestrial life.

The extraterrestric life is mimicked in laboratories on Earth with many assumptions – perhaps this overview stimulates some thinking. The discussion presented here should be taken as concept about simple replicating and evolving entities possibly arising from different building blocks in other environments, with structure being more relevant than sequence.

AUTHOR CONTRIBUTIONS

Both authors wrote the manuscript and approved the final version.

FUNDING

FB was financially supported by a fellowship of the German Academy of Sciences Leopoldina.

ACKNOWLEDGMENTS

Dedicated in memoriam to Manfred Eigen (1927–2019) with gratitude for invitations to KM to many inspiring Winter Seminars in Klosters, Switzerland. We would also like to thank Prof. Peter Palese for his generous support.

REFERENCES

AbouHaidar, M. G., Venkataraman, S., Golshani, A., Liu, B., and Ahmad, T. (2014). Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14542–14547. doi: 10.1073/pnas.1402814111

Altman, S. (2013). The RNA-protein-world. *RNA* 19, 589–590. doi: 10.1261/rna.038687.113

Andras, P., and Andras, C. (2005). The origin of life - the 'protein interaction world' hypothesis: protein interactions were the first form of self-reproducing life and nucleic acids evolved later as memory molecules. *Med. Hypotheses* 64, 678–688. doi: 10.1016/j.mehy.2004.11.029

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361:eaar7191. doi: 10.1126/science.aar7191

Attwater, J., Wochner, A., Pinheiro, V. B., Coulson, A., and Holliger, P. (2010). Ice as a protocellular medium for RNA replication. *Nat. Commun.* 1:76. doi: 10.1038/ncomms1076

Beare, P. A., Samuel, J. E., Howe, D., Virtaneva, K., Porcella, S. F., and Heinzen, R. A. (2006). Genetic diversity of the Q fever agent, Coxiella burnetii, assessed by microarray-based whole-genome comparisons. *J. Bacteriol.* 188, 2309–2324. doi: 10.1128/JB.188.7.2309-2324.2006

Béguin, P., Gill, S., Charpin, N., and Forterre, P. (2015). Synergistic template-free synthesis of dsDNA by Thermococcus nautili primase PolpTN2, DNA polymerase PolB, and pTN2 helicase. *Extremophiles* 19, 69–76. doi: 10.1007/s00792-014-0706-1

Beltz, L. A. (2018). *Bats and Human Health: Ebola, SARS, Rabies and Beyond*. Hoboken, NJ: Wiley-Blackwell.

Brackett, D. M., and Dieckmann, T. (2006). Aptamer to ribozyme: the intrinsic catalytic potential of a small RNA. *Chembiochem* 7, 839–843. doi: 10.1002/cbic.200500538

Broecker, F., Hardt, C., Herwig, R., Timmermann, B., Kerick, M., Wunderlich, A., et al. (2016a). Transcriptional signature induced by a metastasis-promoting c-Src mutant in a human breast cell line. *FEBS J.* 283, 1669–1688. doi: 10.1111/febs.13694

Broecker, F., Horton, R., Heinrich, J., Franz, A., Schweiger, M. R., Lehrach, H., et al. (2016b). The intron-enriched HERV-K(HML-10) family suppresses apoptosis, an indicator of malignant transformation. *Mob. DNA* 7:25. doi: 10.1186/s13100-016-0081-9

Broecker, F., Klumpp, J., and Moelling, K. (2016c). Long-term microbiota and virome in a Zürich patient after fecal transplantation against Clostridium difficile infection. *Ann. N. Y. Acad. Sci.* 1372, 29–41. doi: 10.1111/nyas.13100

Broecker, F., Russo, G., Klumpp, J., and Moelling, K. (2017). Stable core virome despite variable microbiome after fecal transfer. *Gut Microbes* 8, 214–220. doi: 10.1080/19490976.2016.1265196

Cech, T. R. (2000). Structural biology. The ribosome is a ribozyme. *Science* 289, 878–879.

Charpentier, E., and Doudna, J. A. (2013). Biotechnology: rewriting a genome. *Nature* 495, 50–51. doi: 10.1038/495050a

Chela-Flores, J. (1994). Are viroids molecular fossils of the RNA world? *J. Theor. Biol.* 166, 163–166. doi: 10.1006/jtbi.1994.1014

Colson, P., Levasseur, A., La Scola, B., Sharma, V., Nasir, A., Pontarotti, P., et al. (2018). Ancestrality and mosaicism of giant viruses supporting the definition of the fourth TRUC of microbes. *Front. Microbiol.* 9:2668. doi: 10.3389/fmicb.2018.02668

Cordaux, R., and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.* 10, 691–703. doi: 10.1038/nrg2640

Cotillard, A., Kennedy, S. P., Kong, L. C., Prifti, E., Pons, N., Le Chatelier, E., et al. (2013). Dietary intervention impact on gut microbial gene richness. *Nature* 500, 585–588. doi: 10.1038/nature12480

Cotton, J. A., and Page, R. D. (2005). Rates and patterns of gene duplication and loss in the human genome. *Proc. Biol. Sci.* 272, 277–283. doi: 10.1098/rspb.2004.2969

Crick, F. H. (1968). The origin of the genetic code. *J. Mol. Biol.* 38, 367–379. doi: 10.1016/0022-2836(68)90392-6

de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7:e1002384. doi: 10.1371/journal.pgen.1002384

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat. Rev. Genet.* 6, 361–375. doi: 10.1038/nrg1603

Deveson, I. W., Hardwick, S. A., Mercer, T. R., and Mattick, J. S. (2017). The dimensions, dynamics, and relevance of the mammalian noncoding transcriptome. *Trends Genet.* 33, 464–478. doi: 10.1016/j.tig.2017.04.004

Dewannieux, M., and Heidmann, T. (2013). Endogenous retroviruses: acquisition, amplification and taming of genome invaders. *Curr. Opin. Virol.* 3, 646–656. doi: 10.1016/j.coviro.2013.08.005

Diener, T. O. (2016). Viroids: "living fossils" of primordial RNAs? *Biol. Direct* 11:15. doi: 10.1186/s13062-016-0116-7

Dreher, T. W. (2009). Role of tRNA-like structures in controlling plant virus replication. *Virus Res.* 139, 217–229. doi: 10.1016/j.virusres.2008.06.010

Dreher, T. W. (2010). Viral tRNAs and tRNA-like structures. *Wiley Interdiscip. Rev. RNA* 1, 402–414. doi: 10.1002/wrna.42

Dupressoir, A., Lavialle, C., and Heidmann, T. (2012). From ancestral infectious retroviruses to bona fide cellular genes: role of the captured syncytins in placentation. *Placenta* 33, 663–671. doi: 10.1016/j.placenta.2012.05.005

Eigen, M. (1971). Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58, 465–523. doi: 10.1007/BF00623322

Eigen, M. (2013). *From Strange Simplicity to Complex Familiarity*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198570219.001.0001

Embley, T. M., and Martin, W. (2006). Eukaryotic evolution, changes and challenges. *Nature* 440, 623–630. doi: 10.1038/nature04546

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296. doi: 10.1038/nrg3199

Fischer, M. G., and Hackl, T. (2016). Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* 540, 288–291. doi: 10.1038/nature20593

Flint, S. J. (2015). *Principles of Virology*, 4th Edn. Washington, DC: ASM Press. doi: 10.1128/9781555819521

Flores, R., Gago-Zachert, S., Serra, P., Sanjuán, R., and Elena, S. F. (2014). Viroids: survivors from the RNA world? *Annu. Rev. Microbiol.* 68, 395–414. doi: 10.1146/annurev-micro-091313-103416

Flores, R., Hernández, C., Martínez de Alba, A. E., Daròs, J. A., and Di Serio, F. (2005). Viroids and viroid-host interactions. *Annu. Rev. Phytopathol.* 43, 117–139. doi: 10.1146/annurev.phyto.43.040204.140243

Galvani, A. P., and Slatkin, M. (2003). Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15276–15279. doi: 10.1073/pnas.2435085100

Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., et al. (2008). Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* 319, 1215–1220. doi: 10.1126/science.1151721

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi: 10.1126/science.1190719

Gilbert, W. (1986). Origin of life: the RNA world. *Nature* 319:618. doi: 10.1038/319618a0

Giovannoni, S. J. (2017). SAR11 bacteria: the most abundant plankton in the Oceans. *Ann. Rev. Mar. Sci.* 9, 231–255. doi: 10.1146/annurev-marine-010814-015934

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309, 1242–1245. doi: 10.1126/science.1114057

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442, 199–202. doi: 10.1038/nature04917

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522, 221–225. doi: 10.1038/nature14308

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Harish, A., and Caetano-Anolles, G. (2012). Ribosomal history reveals origins of modern protein synthesis. *PLoS One* 7:e32776. doi: 10.1371/journal.pone.0032776

Hayward, A. (2017). Origin of retroviruses: when, where, and how? *Curr. Opin. Virol.* 25, 23–27. doi: 10.1016/j.coviro.2017.06.006

Holmes, E. C. (2011). What does virus evolution tell us about virus origins? *J. Virol.* 85, 5247–5251. doi: 10.1128/JVI.02203-10

Huber, H., Hohn, M. J., Stetter, K. O., and Rachel, R. (2003). The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life. *Res. Microbiol.* 154, 165–171. doi: 10.1016/S0923-2508(03)00035-4

Hutchison, C. A. III, Chuang, R. Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., et al. (2016). Design and synthesis of a minimal bacterial genome. *Science* 351:aad6253. doi: 10.1126/science.aad6253

Joyce, G. F. (2002). The antiquity of RNA-based evolution. *Nature* 418, 214–221. doi: 10.1038/418214a

Kacian, D. L., Mills, D. R., Kramer, F. R., and Spiegelman, S. (1972). A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. *Proc. Natl. Acad. Sci. U.S.A.* 69, 3038–3042. doi: 10.1073/pnas.69.10.3038

Kasting, J. F., Whitmire, D. P., and Reynolds, R. T. (1993). Habitable zones around main sequence stars. *Icarus* 101, 108–128. doi: 10.1006/icar.1993.1010

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Koskiniemi, S., Sun, S., Berg, O. G., and Andersson, D. I. (2012). Selection-driven gene loss in bacteria. *PLoS Genet.* 8:e1002787. doi: 10.1371/journal.pgen.1002787

Kreysing, M., Keil, L., Lanzmich, S., and Braun, D. (2015). Heat flux across an open pore enables the continuous replication and selection of oligonucleotides towards increasing length. *Nat. Chem.* 7, 203–208. doi: 10.1038/nchem.2155

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., et al. (2003). A giant virus in amoebae. *Science* 299:2033. doi: 10.1126/science.1081867

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., et al. (2008). The virophage as a unique parasite of the giant mimivirus. *Nature* 455, 100–104. doi: 10.1038/nature07218

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature* 500, 541–546. doi: 10.1038/nature12506

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4274–4279. doi: 10.1073/pnas.1320670111

Lescot, M., Hingamp, P., Kojima, K. K., Villar, E., Romac, S., Veluchamy, A., et al. (2016). Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. *ISME J.* 10, 1134–1146. doi: 10.1038/ismej.2015.192

Li, H., Li, L., Liu, L. R., Omange, R. W., Toledo, N., Kashem, M. A., et al. (2018). Hypothetical endogenous SIV-like antigens in Mauritian cynomolgus macaques. *Bioinformation* 14, 48–52. doi: 10.6026/97320630014048

Li, H., Nykoluk, M., Li, L., Liu, L. R., Omange, R. W., Soule, G., et al. (2017). Natural and cross-inducible anti-SIV antibodies in Mauritian cynomolgus macaques. *PLoS One* 12:e0186079. doi: 10.1371/journal.pone.0186079

Lincoln, T. A., and Joyce, G. F. (2009). Self-sustained replication of an RNA enzyme. *Science* 323, 1229–1232. doi: 10.1126/science.1167856

Ma, B. G., Chen, L., Ji, H. F., Chen, Z. H., Yang, F. R., Wang, L., et al. (2008). Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* 366, 607–611. doi: 10.1016/j.bbrc.2007.12.014

Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K., et al. (2014). The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* 42, 4160–4179. doi: 10.1093/nar/gkt1414

Malik, H. S., and Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res.* 11, 1187–1197. doi: 10.1101/gr.185101

Margulis, L. (1970). *Origin of Eukaryotic Cells*. New Haven, CT: Yale University Press.

Margulis, L. (1993). *Symbiosis in Cell Evolution*, 2nd Edn. New York, NY: W.H. Freeman and Co.

Martin, W., Baross, J., Kelley, D., and Russell, M. J. (2008). Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* 6, 805–814. doi: 10.1038/nrmicro1991

Meierhenrich, U. (2008). *Amino Acids and the Asymmetry of Life*. Berlin: Springer-Verlag.

Miller, S. L. (1953). A production of amino acids under possible primitive earth conditions. *Science* 117, 528–529. doi: 10.1126/science.117.3046.528

Miller, S. L., and Urey, H. C. (1959). Organic compound synthesis on the primitive Earth. *Science* 130, 245–251. doi: 10.1126/science.130.3370.245

Mitch, L. (2017). Giant viruses found in Austrian sewage fuel debate over potential fourth domain of life. *Science* doi: 10.1126/science.aal1005

Moelling, K. (2012). Are viruses our oldest ancestors? *EMBO Rep.* 13:1033. doi: 10.1038/embor.2012.173

Moelling, K. (2013). What contemporary viruses tell us about evolution: a personal view. *Arch. Virol.* 158, 1833–1848. doi: 10.1007/s00705-013-1679-6

Moelling, K. (2017). *Viruses, More Friends than Foes*. Toh Tuck: World Scientific Press.

Moelling, K., and Broecker, F. (2015). The reverse transcriptase-RNase H: from viruses to antiviral defense. *Ann. N. Y. Acad. Sci.* 1341, 126–135. doi: 10.1111/nyas.12668

Moelling, K., Broecker, F., Russo, G., and Sunagawa, S. (2017). RNase H as gene modifier, driver of evolution and antiviral defense. *Front. Microbiol.* 8:1745. doi: 10.3389/fmicb.2017.01745

Moelling, K., Heimann, B., Beimling, P., Rapp, U. R., and Sander, T. (1984). Serine- and threonine-specific protein kinase activities of purified gag-mil and gag-raf proteins. *Nature* 312, 558–561. doi: 10.1038/312558a0

Moelling, K., Matskevich, A., and Jung, J. S. (2006). Relationship between retroviral replication and RNA interference machineries. *Cold Spring Harb. Symp. Quant. Biol.* 71, 365–368. doi: 10.1101/sqb.2006.71.010

Mölling, K., Bolognesi, D. P., Bauer, H., Büsen, W., Plassmann, H. W., and Hausen, P. (1971). Association of viral reverse transcriptase with an enzyme degrading the RNA moiety of RNA-DNA hybrids. *Nat. New Biol.* 234, 240–243. doi: 10.1038/newbio234240a0

Müller, G., Strack, B., Dannull, J., Sproat, B. S., Surovoy, A., Jung, G., et al. (1994). Amino acid requirements of the nucleocapsid protein of HIV-1 for increasing catalytic activity of a Ki-ras ribozyme in vitro. *J. Mol. Biol.* 242, 422–429. doi: 10.1006/jmbi.1994.1592

Nasir, A., and Caetano-Anolles, G. (2015). A phylogenomic data-driven exploration of viral origins and evolution. *Sci. Adv.* 1:e1500527. doi: 10.1126/sciadv.1500527

Ogata, H., Audic, S., Renesto-Audiffren, P., Fournier, P. E., Barbe, V., Samson, D., et al. (2001). Mechanisms of evolution in Rickettsia conorii and *R. prowazekii*. *Science* 293, 2093–2098.

Orgel, L. E. (2004). Prebiotic chemistry and the origin of the RNA world. *Crit. Rev. Biochem. Mol. Biol.* 39, 99–123. doi: 10.1080/10409230490460765

Ott, G., Rosenwald, A., and Campo, E. (2013). Understanding MYC-driven aggressive B-cell lymphomas: pathogenesis and classification. *Blood* 122, 3884–3891. doi: 10.1182/blood-2013-05-498329

Ott, S. J., Waetzig, G. H., Rehman, A., Moltzau-Anderson, J., Bharti, R., Grasis, J. A., et al. (2017). Efficacy of sterile fecal filtrate transfer for treating patients with clostridium difficile infection. *Gastroenterology* 152, 799–811.e7. doi: 10.1053/j.gastro.2016.11.010

Patel, B. H., Percivalle, C., Ritson, D. J., Duffy, C. D., and Sutherland, J. D. (2015). Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* 7, 301–307. doi: 10.1038/nchem.2202

Pennazio, S., and Roggero, P. (2000). The discovery of the chemical nature of tobacco mosaic virus. *Riv. Biol.* 93, 253–281.

Philippe, N., Legendre, M., Doutre, G., Couté, 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. doi: 10.1126/science.1239181

Riesner, D., Henco, K., Rokohl, U., Klotz, G., Kleinschmidt, A. K., Domdey, H., et al. (1979). Structure and structure formation of viroids. *J. Mol. Biol.* 133, 85–115. doi: 10.1016/0022-2836(79)90252-3

Roberts, R. W., and Szostak, J. W. (1997). RNA-peptide fusione for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U.S.A.* 94, 12297–12302. doi: 10.1073/pnas.94.23.12297

Schmalzbauer, E., Strack, B., Dannull, J., Guehmann, S., and Moelling, K. (1996). Mutations of basic amino acids of NCp7 of human immunodeficiency virus type 1 affect RNA binding in vitro. *J. Virol.* 70, 771–777.

Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., et al. (2017). Giant viruses with an expanded complement of translation system components. *Science* 356, 82–85. doi: 10.1126/science.aal4657

Segata, N. (2015). Gut microbiome. Westernization and the disappearance of intestinal diversity. *Curr. Biol.* 25, R611–R613. doi: 10.1016/j.cub.2015.05.040

Segré, D., Ben-Eli, D., Deamer, D. W., and Lancet, D. (2001). The lipid world. *Orig. Life Evol. Biosph.* 31, 119–145. doi: 10.1023/A:1006746807104

Simon, D. M., and Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res.* 36, 7219–7229. doi: 10.1093/nar/gkn867

Solloch, U. V., Lang, K., Lange, V., Böhme, I., Schmidt, A. H., and Sauter, J. (2017). Frequencies of gene variant CCR5-Δ32 in 87 countries based on next-generation sequencing of 1.3 million individuals sampled from 3 national DKMS donor centers. *Hum. Immunol.* 78, 710–717. doi: 10.1016/j.humimm.2017.10.001

Spiegelman, S., Haruna, I., Holland, I. B., Beaudreau, G., and Mills, D. (1965). The synthesis of a self-propagating and infectious nucleic acid with a purified enzyme. *Proc. Natl. Acad. Sci. U.S.A.* 54, 919–927. doi: 10.1073/pnas.54.3.919

Steele, E. J., Al-Mufti, S., Augustyn, K. A., Chandrajith, R., Coghlan, J. P., Coulson, S. G., et al. (2018). Cause of cambrian explosion - terrestrial or cosmic? *Prog. Biophys. Mol. Biol.* 136, 3–23. doi: 10.1016/j.pbiomolbio.2018.03.004

Sumper, M., and Luce, R. (1975). Evidence for de novo production of self-replicating and environmentally adapted RNA structures by bacteriophage Qbeta replicase. *Proc. Natl. Acad. Sci. U.S.A.* 72, 162–166. doi: 10.1073/pnas.72.1.162

Suttle, C. A. (2005). Viruses in the sea. *Nature* 437, 356–361. doi: 10.1038/nature04160

Szostak, J. W. (2011). An optimal degree of physical and chemical heterogeneity for the origin of life? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 366, 2894–2901. doi: 10.1098/rstb.2011.0140

Szostak, J. W., Bartel, D. P., and Luisi, P. L. (2001). Synthesizing life. *Nature* 409, 387–390. doi: 10.1038/35053176

Tarlinton, R. E. (2012). "Koala retrovirus endogenisation in action," in *Viruses: Essential Agents of Life*, ed. G. Witzany (Berlin: Springer), 283–291. doi: 10.1007/978-94-007-4899-6_14

Taylor, J. M. (2009). Replication of the hepatitis delta virus RNA genome. *Adv. Virus Res.* 74, 103–121. doi: 10.1016/S0065-3527(09)74003-5

Temin, H. M. (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Mol. Biol. Evol.* 2, 455–468. doi: 10.1093/oxfordjournals.molbev.a040365

Vasas, V., Szathmary, E., and Santosa, M. (2010). Lack of evolvability in self-sustaining autocatalytic networks: a constraint on the metabolism-fist path to the origin of life. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1470–1475. doi: 10.1073/pnas.0912628107

Villarreal, L. P. (2009). The source of self: genetic parasites and the origin of adaptive immunity. *Ann. N. Y. Acad. Sci.* 1178, 194–232. doi: 10.1111/j.1749-6632.2009.05020.x

Villarreal, L. P. (2011). Viral ancestors of antiviral systems. *Viruses* 3, 1933–1958. doi: 10.3390/v3101933

Villarreal, L. P., and Witzany, G. (2010). Viruses are essential agents within the roots and stem of the tree of life. *J. Theor. Biol.* 262, 698–710. doi: 10.1016/j.jtbi.2009.10.014

Vogelstein, B., and Kinzler, K. W. (2002). *The Genetic Basis of Human Cancer*, 2nd Edn. New York, NY: McGraw-Hill, 116.

Wang, D., Li, L. B., Hou, Z. W., Kang, X. J., Xie, Q. D., Yu, X. J., et al. (2011). The integrated HIV-1 provirus in patient sperm chromosome and its transfer into the early embryo by fertilization. *PLoS One* 6:e28586. doi: 10.1371/journal.pone.0028586

Wilson, D. S., and Szostak, J. W. (1999). In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.* 68, 611–647. doi: 10.1146/annurev.biochem.68.1.611

Wolf, Y. I., and Koonin, E. V. (2013). Genome reduction as the dominant mode of evolution. *Bioessays* 35, 829–837. doi: 10.1002/bies.201300037

Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* 9, 3353–3362. doi: 10.1002/j.1460-2075.1990.tb07536.x

Yang, W. T., Wu, L. W., Tseng, T. C., Chen, C. L., Yang, H. C., Su, T. H., et al. (2016). Hepatitis B surface antigen loss and hepatocellular carcinoma development in patients with dual hepatitis B and C infection. *Medicine* 95:e2995. doi: 10.1097/MD.0000000000002995

Zhang, B., and Cech, T. R. (1997). Peptide bond formation by in vitro selected ribozymes. *Nature* 390, 96–100. doi: 10.1038/36375

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18, 292–298. doi: 10.1016/S0169-5347(03)00033-8

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., et al. (2013). Abundant SAR11 viruses in the ocean. *Nature* 494, 357–360. doi: 10.1038/nature11921

Conflict of Interest Statement: 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.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 29 March 2019
doi: 10.3389/fmicb.2019.00584

Type II Restriction of Bacteriophage DNA With 5hmdU-Derived Base Modifications

Kiersten Flodman[†], Rebecca Tsai[†], Michael Y. Xu[†], Ivan R. Corrêa Jr., Alyssa Copelas, Yan-Jiun Lee, Ming-Qun Xu, Peter Weigele and Shuang-yong Xu*

New England Biolabs, Inc., Ipswich, MA, United States

To counteract bacterial defense systems, bacteriophages (phages) make extensive base modifications (substitutions) to block endonuclease restriction. Here we evaluated Type II restriction of three thymidine (T or 5-methyldeoxyuridine, 5mdU) modified phage genomes: *Pseudomonas* phage M6 with 5-(2-aminoethyl)deoxyuridine (5-*N*edU), *Salmonella* phage ViI (Vi1) with 5-(2-aminoethoxy)methyldeoxyuridine (5-*N*e*O*mdU) and *Delftia* phage phi W-14 (a.k.a. ΦW-14) with α-putrescinylthymidine (putT). Among >200 commercially available restriction endonucleases (REases) tested, phage M6, ViI, and phi W-14 genomic DNAs (gDNA) show resistance against 48.4, 71.0, and 68.8% of Type II restrictions, respectively. Inspection of the resistant sites indicates the presence of conserved dinucleotide TG or TC (TS, S=C, or G), implicating the specificity of TS sequence as the target that is converted to modified base in the genomes. We also tested a number of DNA methyltransferases (MTases) on these phage DNAs and found some MTases can fully or partially modify the DNA to confer more resistance to cleavage by REases. Phage M6 restriction fragments can be efficiently ligated by T4 DNA ligase. Phi W-14 restriction fragments show apparent reduced rate in *E. coli* exonuclease III degradation. This work extends previous studies that hypermodified T derived from 5hmdU provides additional resistance to host-encoded restrictions, in parallel to modified cytosines, guanine, and adenine in phage genomes. The results reported here provide a general guidance to use REases to map and clone phage DNA with hypermodified thymidine.

Keywords: Type II restriction and modification, 5hmdU-derived nucleotide modification, *Pseudomonas* bacteriophage (phage) M6, *Salmonella* phage ViI, *Delftia* phi W-14, phage therapy

OPEN ACCESS

Edited by:
Arshan Nasir,
COMSATS University Islamabad,
Pakistan

Reviewed by:
Bob Gordon Blasdel,
Vesale Pharma, Belgium
Andrei A. Zimin,
Institute of Biochemistry
and Physiology of Microorganisms
(RAS), Russia

****Correspondence:***
Shuang-yong Xu
xus@neb.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: *27 January 2019*
Accepted: *07 March 2019*
Published: *29 March 2019*

Citation:
Flodman K, Tsai R, Xu MY,
Corrêa IR Jr, Copelas A, Lee Y-J,
Xu M-Q, Weigele P and Xu S-y (2019)
Type II Restriction of Bacteriophage
DNA With 5hmdU-Derived Base
Modifications.
Front. Microbiol. 10:584.
doi: 10.3389/fmicb.2019.00584

INTRODUCTION

Type II restriction and modification (R-M) systems in bacteria encode restriction endonucleases (REases) to destroy invading foreign DNA in phage infection and acquisition of mobile genetic elements (Smith and Wilcox, 1970; Landy et al., 1974, reviewed in Pingoud et al., 2016). To gain an upper hand in the biowarfare, bacteriophages (phages) utilize DNA base modifications [or nucleotide (nt) substitutions] to counteract host-encoded Type II restrictions (Huang et al., 1982; Kruger and Bickle, 1983; Miller et al., 1985; Kulikov et al., 2014; Tsai et al., 2017; Lee et al., 2018). Extensive non-canonical nt substitutions have been reported for all four bases in DNA: for example, 5-methylcytosine (5mC) replacing all C in phage XP12 genome (Feng et al., 1978), 5-glucosylated-hydroxymethylcytosines (5gmC) in phage T4 (Gold and Schweiger, 1969), deoxyarchaeosine (dG^+) in phage 9g (Thiaville et al., 2016; Tsai et al., 2017), α-putrescinylthymidine (putT) in phi W-14 (Kelln and Warren, 1973; Kropinski et al., 1973), *N*6-methyladenine (6mA) in some phages encoding frequent adenine methyltranferases (MTases) (Drozdz et al., 2012; Murray et al., 2018), and 5-hydroxymethyluridine (5hmdU) in *Bacillus* phage SP8 and SPO1 (Stewart et al., 2009, reviewed in Weigele and Raleigh, 2016). Non-canonical nt substitutions can be introduced during DNA replication through modified dNTP (e.g., 5hmCTP, 5hmdUTP). Further base modification can be carried out post-replicationally by phage-encoded enzymes such as MTases, DNA glycosyltranferases, and alkylamine transferases. In some cases, phage DNA is partially modified by host-encoded enzymes, such as Dcm (methylation of CCWGG to C5mCWGG) and Dam methyltransferases (methylation of GATC to G6mATC), or other MTases during phage DNA replication. Phage T4 DNA modified with 5gmC is resistant to many Type II REases that recognize GC-containing sequences (Huang et al., 1982). Phage 9g DNA with the dG^+ modification is resistant to ∼71% of Type II restriction specificities with GC sequences (Tsai et al., 2017). In a limited Type II restriction study of phage phi W-14 genome containing putT, 17 out of 32 REases tested were blocked by the base modification (Miller et al., 1985). The additional positive charges from the side chain of putT likely interfere with restriction enzyme tracking process due to altered local DNA structure since the cleavage efficiency of some REases with only GC sequence was also impaired. Phi W-14 genomic DNA is packaged more compactly in phage head than T4 (Scraba et al., 1983). Phage SPO1 genomic DNA with 5hmdU substitutions is fully resistant to 4 out of 30 Type II restrictions (∼13.3%) and partially resistant (i.e., slower cleavage in 1 h digestion) to 17 out of 30 REases tested (56.7%) (Huang et al., 1982; Vilpo and Vilpo, 1995). DNA duplex oligos with the 5hmdU substitution display reduced melting temperature (Tm) and altered backbone flexibility when passing through nanopores (Carson et al., 2016).

Two new base modifications, 5-(2-aminoethyl)deoxyuridine (5-*N*edU) and 5-(2-aminoethoxy)methyldeoxyuridine (5-*N*eOmdU) were recently discovered in the genomes of *Pseudomonas* phage M6 and *Salmonella* phage ViI (Vi1) (Lee et al., 2018). Hypermodified *Pseudomonas* phage DNAs were shown to be resistant to Type II restriction. Phage M6 and ViI encode a modification gene cluster in their genomes for the production of 5hmdU and the enzymes responsible for subsequent reactions to add the desired chemical groups (Lee et al., 2018). It has been predicted that these phages also encode their own primase, DNA polymerase/clamp loader protein/sliding clamp holder protein, DNA ligase, and RNase H, all of which displaying specialized properties to incorporate modified dNTP intermediate during replication. The three phages M6, ViI, and phi W14 containing hypermodified thymidine bases are thought to utilize the common intermediate 5hmdU. 5hmdU is incorporated into DNA, then phosphorylated by a 5hmdU DNA kinase, and further modified by alkylamine transferases and other associated enzymes. Not all thymidines in the genome are replaced by 5hmdU; in addition to the hypermodified base, these phage DNAs may also carry regular base T and 5hmdU. Bioinformatic prediction of enzymes involved in phage nucleotide hypermodifications has provided abundant information on gene clusters and biosynthetic pathways (Iyer et al., 2013).

The goal of this work is to examine Type II restrictions of modified DNA in phage M6, ViI, and phi W-14 genomes. We performed restriction digestions of these three gDNAs to verify their resistant level *in vitro*. We also analyzed the resistant sites for any conserved sequence motifs to shed light on possible modification site specificity. Furthermore, we introduced additional base modifications in their DNA by treatment with cytosine or adenine MTases to generate two types of base modifications (for instance in M6 DNA, a combination of 5mC and 5-*N*edU, or 6mA and 5-*N*edU). We also examined the ligation efficiency of phage DNA restriction fragments and tested two exonuclease activity on the modified DNA. This work provides basic information on restriction of T-modified DNA and further our understanding of the co-evolution relationship of host and hypermodified phage genomes. Study of highly modified phage genomes may have impact in phage therapy.

MATERIALS AND METHODS

Phage DNA Purification and Restriction Digestions

REases, MTases, DNA ligase, DNA nuclease, and phosphatase, Proteinase K, exonuclease, and repair enzyme hSMUG1 were provided by New England Biolabs (NEB). Phage particles were purified by CsCl gradient method and phage DNA purified by phenol-$CHCl_3$ extraction, and ethanol precipitation (Sambrook et al., 1989). Due to poor phage titer of M6 phage, phage infection and propagation were carried out on solid growth medium and phage lysates were pooled from multiple plates. NEBcutter V2.1 software (Vincze et al., 2003) was used to generate restriction patterns of phage DNA with the assumption of no base modification. We used excess of REases in restriction digestions (5 to 40 U to cleave 0.25 to 0.5 μg phage DNA) in 50 μl total volume incubated at the recommended temperature for 1 h (e.g., 5 μl of REases for low concentration enzyme supplied at 1000 U/ml, 2 μl of REase for high concentration REase supplied at 20,000 U/ml). Digested DNAs were analyzed by agarose gel electrophoresis. The DNA cleavage patterns were compared to NEBcutter-generated restriction patterns to determine digestion results as complete (c), partial (p), very partial (vp), or resistant (x) to digestions. For digestion of viral DNA with glycosylase and AP endonuclease, DNA was first incubated with hSMUG1 for 1 h, and then treated with *Escherichia coli* endonuclease VIII.

Methylation and Challenge With REases to Check Methylation Level

Phage DNA was methylated by treatment with excess DNA MTase and methyl-donor SAM in the recommended buffer for 2 h. Following Proteinase K treatment and spin column purification, the methylated DNA was digested by cognate or non-cognate REases to evaluate the degree of resistance to restriction.

Methylation and Determination of Base Compositions by Liquid Chromatography-Mass Spectrometry (LC-MS)

Phage DNA was methylated by the frequent MTases M.EcoGII (adenine methyltransferase), M.SssI (CpG methyltransferase), M.CviPI (GpC methyltransferase) for 2–4 h with methyl donor SAM. After Proteinase K treatment, the DNA was precipitated in ethanol, dried and resuspended in a buffer for nuclease degradation. DNA samples (5 μg) were digested to nucleosides by treatment with the Nucleoside Digestion Mix (NEB, M0649S) overnight at 37°C. Nucleoside analysis was performed on an Agilent LC/MS System 1200 Series instrument equipped with a G1315D diode array detector and a 6120 Single Quadrupole Mass Detector operating in positive (+ESI) and negative (−ESI) electrospray ionization modes. LC was carried out on a Waters Atlantis T3 column (4.6 mm × 150 mm, 3 μm) with a gradient mobile phase consisting of 10 mM aqueous ammonium acetate (pH 4.5) and methanol. MS data acquisition was recorded in total ion chromatogram (TIC) mode. Each nucleoside was identified as follows: dC $[M + H]^+$ 228.1 and $[M\text{-}H]^-$ 226.2; dG $[M + H]^+$ 268.1 and $[M\text{-}H]^-$ 266.1; dT $[M + H]^+$ 243.1 and $[M\text{-}H]^-$ 241.1; dA $[M + H]^+$ 252.1 and $[M\text{-}H]^-$ 250.1; 5mdC $[M + H]^+$ 242.1 and $[M\text{-}H]^-$ 240.2; 6mdA $[M + H]^+$ 266.1 and $[M\text{-}H]^-$ 264.1; 5hmdC $[M\text{-}H]^-$ 257.0; 5-*NeO*mdU $[M + H]^+$ 302.1 and $[M\text{-}H]^-$ 300.1; α-putT $[M + H]^+$ 329.2 and $[M\text{-}H]^-$ 327.2; putT-dC $[M + H]^+$ 618.3 and $[M\text{-}H]^-$ 616.2; and putT-dG $[M + H]^+$ 658.2 and $[M\text{-}H]^-$ 656.1. The relative abundance of each nucleoside was determined by dividing the UV absorbance by the corresponding extinction coefficient at 260 nm.

RESULTS

Restriction of Phage M6, ViI, and phi W-14 Genomic DNA

To find out the resistance level, we carried out restriction digestions for phage M6, ViI, and phi W-14 genomic DNA. The chemical structure of the modified bases discussed in this work is shown in **Supplementary Figure 1**. It was unknown beforehand how many units are required for complete digestion of each modified DNA since the unit definition is typically done on phage λ or pBR322 DNA by the manufacturer. We used phage λ and pTYB2 DNA for control digestions to validate REases that are active, but not able to cleave modified DNA. The restriction of modified phage DNA was repeated at least once to confirm reproducibility. We grouped restriction results into four categories: complete, partial, very partial (most of the substrate DNA remains intact, only a few weak bands visible), and resistant as compared to computer generated banding patterns. The results are shown in **Figures 1A–C** and **Table 1**. Phage M6, ViI, and phi W-14 DNAs are resistant to approximately 48.4, 71.0, and 68.8% of Type II restrictions, as compared to phage 9g DNA (dG^+ modification) resistance to nearly 71% REases tested. The individual restriction results for three genomic DNAs are shown in **Supplementary Tables 1–3**. Phage M6 DNA is completely resistant to *Fsp*I (TGCGCA) and *Sac*I (GAGCTC) restriction, most likely due the modified T in TG and TC dinucleotide in both strands (see below for more resistant site analysis). Phage ViI DNA is resistant to restriction by *Bsp*HI (TCATGA), *Cla*I (ATCGAT), and *Nde*I (CATATG). Phi W-14 DNA is resistant to restriction by *Hpy*188III (TCNNGA) and *Hpy*CH4V (TGCA) probably due to the modified bases in TG or TC dinucleotides in both strands. In some cases, phage DNA is also partially or completely resistant to REases that cleave target sites with 4–6 AT bp (see **Supplementary Tables 1–3**). We concluded that the longer side-chain modifications of phages ViI and phi W-14 DNAs are more effective at blocking Type II restriction than is the smaller aminoethyl group of phage M6 DNA. However, 5-*N*edU shows better resistance than phage DNA with 5hmdU alone (Vilpo and Vilpo, 1995). The partial positive charges of the side chain in the major groove of DNA may affect the indirect read out of target sequence by REases. The phage DNA sensitivity to Type II restriction is also shown in "pie" charts (**Supplementary Figure 2**). Since most of the restriction reactions were carried out with excess enzymes in an overdigestion protocol, we cannot rule out the possibility that some very partial digestions are caused by relaxed "star" activity (restriction enzyme "star" activity can cleave target sites with 1–2 bp off from the canonical sites) (Robinson and Sligar, 1993). Engineered high-fidelity REases were used where available to minimize "star" activity (Vasu et al., 2013). Thus, the resistance level might be underestimated compared to the *in vivo* restriction level. *In vivo* restriction gene expression is tightly regulated by transcription factors such as the C (controller) protein to prevent self-restriction (Tao and Blumenthal, 1992; Sawaya et al., 2013).

There are a number of REases that recognize and cleave target sites with GC bp sequence only. Interestingly, they can cut λ and plasmid (pTYB2) DNA; but are unable to cleave M6 and phi W-14 DNA (**Supplementary Figures 6A,B**). We speculate that these REases are extremely sensitive to the nearby base modifications since the probability of TG dinucleotide 5′ to the *Apa*I (GGGCC/C) and *Psp*OMI (G/GGCCC) sites is only 0.25. Similarly, the probability of TG dinucleotide 5′ to the NarI (GG/CGCC) and PluTI (GGCGC/C) is 0.25. NarI and PluTI partially digested a single site plasmid pTYB2 as two sites are probably required for efficient digestion. This group of enzymes include Type IIE and IIF that requires a secondary site (effector site) and extensive looping and enzyme complex interaction (enzyme dimers or tetramers bound to two sites separated by a certain distance) (Roberts et al., 2003). Phi W-14 genomic DNA is resistant or partially resistant to *Apa*I, *Nae*I (GCC/GGC), *Ngo*MIV (G/CCGGC), *Not*I (GC/GGCCGC), or *Psp*OMI digestion (**Supplementary Figure 6B**). The presence of TG dinucleotides (e.g., tGCCGGC) in the flanking sequence may play a role in the resistance, but it cannot explain all resistant sites.

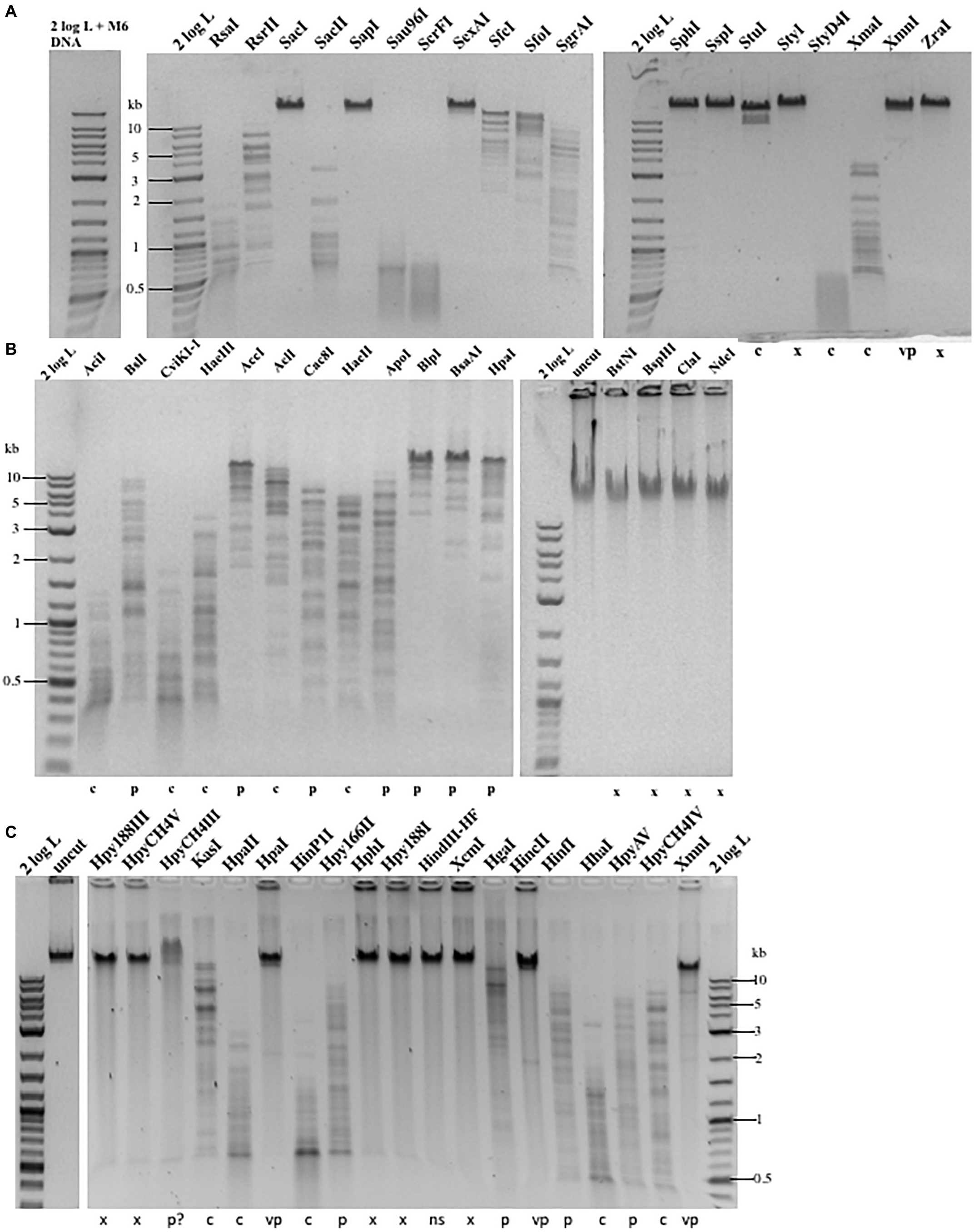


FIGURE 1 | Representative examples of Type II restriction of phage gDNA. **(A)** Phage M6 DNA digested with 19 REases and analyzed by agarose gel electrophoresis. X, resistant to restriction; C, complete digestion; C*, additional fragments observed owing to star activity; P, partial digestion; VP, very partial digestion (only a few weak bands detected); 2 log DNA ladder (0.1–10 kb). Phage M6 DNA is resistant to restriction by *Sac*I (GAGCTC), *Sap*I (GCTCTTC), *Sex*AI (ACCTGGT), *Sph*I (GCATGC), *Ssp*I (AATATT), *Sty*I (CCATGG), *Xmn*I (GAAN4TTC), and *Zra*I (GACGTC), likely due to the presence of TS dinucleotides in one (*Sap*I) or both strands. *Ssp*I site with six Ts in the recognition sequence is also resistant. The restriction results are summarized in **Supplementary Table 1**. The computer-generated restriction patterns by NEBCutter are shown in **Supplementary Figures 3–5**. **(B)** Representative examples of restriction digestions of phage ViI gDNA. The ViI DNA is resistant to restriction by *Bst*NI (CCTGG), *Bsp*HI (TCATGA), *Cla*I (ATCGAT), and *Nde*I (CATATG), likely due to the presence of modified T in TS (TG or TC) dinucleotides in one or both strands. The restriction results are summarized in **Supplementary Table 2**. **(C)** Representative examples of phi W-14 gDNA digested by 19 REases. X, resistant; C, complete digestion; P, partial digestion; VP, very partial digestion; NS, no restriction sites present (as internal negative control); P?, DNA bound and slightly shifted with some smearing; Phi W-14 DNA is resistant to restrictions by *Hpy*188III (TCNNGA), *Hpy*CH4V (TGCA), *Hpa*I (GTTAAC), *Hph*I (GGTGA), *Hpy*188I (TCNGA), *Xcm*I (CCAN9TGG), *Hinc*II (GTYRAC), *Hin*fI (GANTC), and *Xmn*I (GAAN4TTC). The phi W-14 DNA is partially resistant to HpyCH4III (ACNGT) since the probability of having TS sequence with immediate flanking 3′ nt in ACNGTS is 0.5 and the chance of having ACTGT is 0.25. The phage DNA is largely resistant to *Hpa*I (GTTAAC) with tG and 4Ts in both strands. The restriction results are summarized in **Supplementary Tables 1–3**.

TABLE 1 | Type II restriction of phage M6, ViI, and phi W-14 genomic DNA.

Cleavage status	M6	ViI	Phi W-14
Complete	33.7%	8.3%	10.7%
Inconclusive*	1.6%	2.7%	0.9%
Partial	16.3%	18.0%	19.6%
Very partial**	7.9%	6.9%	8.4%
Resistant	40.5%	64.1%	60.4%
Very partial + resistant	48.4%	71.0%	68.8%

**Inconclusive: restriction fragments too large (>10 kb) to be clearly resolved in 0.8–1% agarose gel. **Very partial: most of the genomic DNA remains intact and only a few weak bands were detected.*

Conserved Sequence Motif Among the Resistant Sites in Phage DNA

It has been proposed that M6, ViI, and phi W-14 phages utilize phage-encoded DNA polymerases and a 5hmdUTP, dATP, dCTP, and dGTP deoxynucleotide pool for DNA replication, thus replacing all T with 5hmdU (Neuhard et al., 1980). Further base modifications can occur post-replicationally on the hydroxymethyl moiety of 5hmdU via a phosphorylated intermediate by the action of a 5hmdU DNA kinase (5-HMUDK). It is not known whether the modification site is random or has certain sequence specificity. When the resistant sites were analyzed we observed a predominant sequence motif of TG, TC, TG+TC, or TS+TN dinucleotide. **Table 2** shows that 44 out of 77 resistant sites (57.1%) contain a TG, TC, or TG+TC sequence in phage M6 DNA, while 58.3% of the resistant sites in phage ViI contain the TS motif. The frequency of TS sequence in resistant sites is slightly lower in phage phi W-14 DNA at 49.6%. These numbers are probably underestimated since they do not include the flanking sequence T outside of restriction sites (for example tGCCGGC with a TG dinucleotide). In the restriction analysis of these three phage DNAs, the majority of the resistant sites contain TS dinucleotide in combination with another TT or TA dinucleotide. This suggests that the 5hmdU DNA kinase involved in the phosphorylation of 5hmdU very likely shows the same preference for the TG or TC (TS) sequences. Consistent with the above observation, purified 5hmdU DNA kinase from

TABLE 2 | Frequency of dinucleotide in the resistant sites among the phage genomic DNA.

TN dinucleotides in the resistant sites*			
	Phage M6	Phage ViI	Phi W-14
TG	26	28	29
TC	5	38	24
TC, TG	13	15	15
TG, TA	3	3	3
TG, TT	4	3	4
TG, TN	7	7	10
TC, TT	5	10	6
TC, TA	0	3	3
TC, TN	1	10	9
TT	0	0	0
TA	1	8	7
TN	3	4	8
TT, TA	1	2	1
TT, TN	0	1	0
TA, TN	0	2	3
TC, TG, TT	3	4	4
TA, TC, TN	1	1	2
TA, TT, TN	0	0	3
GC sequence only**	4	?	6
Total	77	139	137

*TN, N = any four nucleotides. TS = TG or TC. *In this resistant site analysis, most of the flanking sequences are excluded. **GC sequence only: a number of REases without A/T in the recognition sequence cannot cleave phage M6 or phi W-14 DNA. But they are active in restriction of λ or plasmid DNA (see examples in **Supplementary Figure 6**). The resistance could be partially due to the 5′ flanking sequence T and negative impact on enzyme tracking process on modified DNA. Highlighted sequences containing TS dinucleotides.*

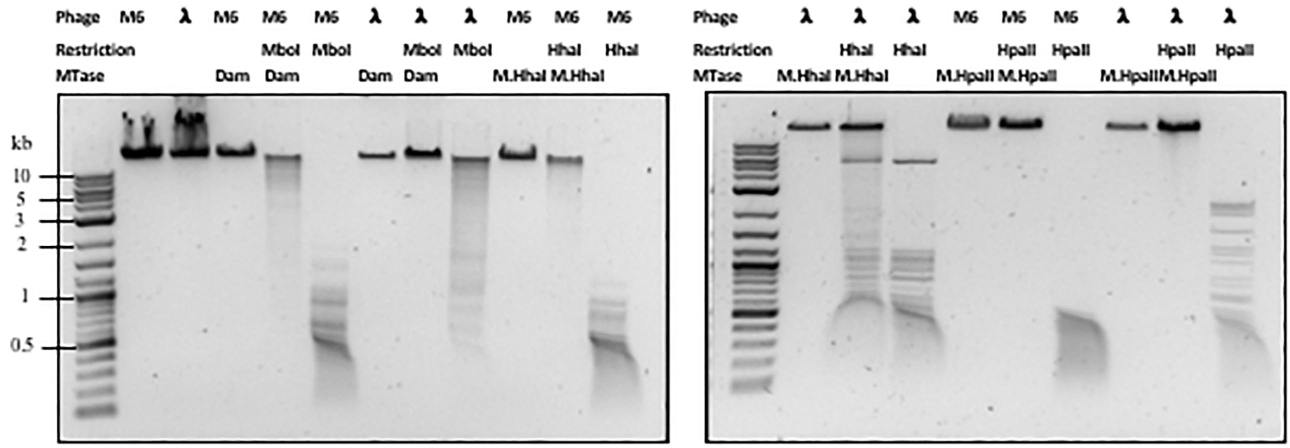


FIGURE 2 | Methylation and restriction of phage M6 gDNA. Phage M6 and λ DNAs were methylated by incubation with indicated MTase and then challenged with cognate or non-cognate REase. Partial restriction digestions resulted from partial modification by the MTase. Resistance to digestions indicate full modification by the MTase.

phage M6 can modify TG dinucleotides in phage SP8 genomic DNA containing 5hmdU (PW, unpublished result).

Methylation of Phage M6 DNA to Generate 5mC- or 6mA-Modified DNA

Next, we examined whether *Pseudomonas* phage M6 DNA can be further modified by C5-cytosine and adenine MTases. This experiment has two possible outcomes: (1) if M6 DNA can be further modified by these MTases, the double-modified DNA may be resistant to more Type II restrictions; (2) 5mC- and N6mA-modified DNA may be subjected to 5mC-dependent restriction systems such as *Eco*K *Mcr*BC (Dila et al., 1990) and *Mcr*A (Mulligan and Dunn, 2008; Czapinska et al., 2018) or N6mA-dependent restriction system *Eco*K Mrr (Heitman and Model, 1987) and Pgl system found in *Streptomyces*, respectively (Hoskisson et al., 2015). Host-acquired modifications of phage genome by Type I MTases were discovered in the early days of molecular biology and phage genetics (Luria and Human, 1952; Bertani and Weigle, 1953; Luria, 1953). Over-expression of a Type II DNA MTase M.BsuM partially modified phage SP10 genome and increased the phage plating efficiency on restriction-proficient (BsuMR$^+$) strain (Matsuoka et al., 2005). In this work we performed DNA methylation and subsequent restriction *in vitro*. After methylation reactions and proteinase K treatment, the phage M6 DNA was purified by spin column and subjected to restriction by the cognate REase or non-cognate endonuclease that is supposed to be blocked by the methylation. **Figure 2** shows representative examples of methylation and restriction experiments. After M.CviPI (GpC methyltransferase) or M.HhaI methylation, the M6 DNA is largely resistant to *Hha*I (GCGC) restriction. M.HpaII and M.MspI can also fully modify M6 DNA and render the DNA resistant to *Hpa*II (CCGG) and *Msp*I (CCGG) restriction, respectively. M.SssI and M.AluI can partially modify the M6 DNA and provide partial resistance to *Hpa*II or *Alu*I (ACGT) restrictions. For the adenine MTases, Dam methyltransferase, M.EcoGII, and M.TaqI can partially modify the M6 DNA and provide partial resistance to *Mbo*I (GATC) and *Taq*I (TCGA) restriction. The methylation and restriction results are summarized in **Table 3**. We concluded that phage M6 DNA can be further modified by C5-cytosine or adenine MTases, which provide additional protection against Type II restriction. The secondary nt modifications might be beneficial for using *Pseudomonas* lytic phages to combat multi-drug resistant *Pseudomonas* infection.

Base Composition Analysis of Methylated Phages ViI and phi W-14 DNA

To estimate the level of methylation in phage ViI and phi W-14 DNAs, we performed LC-MS analysis of the corresponding MTase-treated DNAs. **Figure 3** shows that ∼28% of adenosines have been methylated to 6mA in M.EcoGII-treated ViI genomic DNA. M.CviPI-treated ViI DNA gave rise to ∼7% of 5mC. The composition of the naturally occurring 5-*NeO*mdU, 5-hmdU, and T in phage ViI genomic DNA were estimated at 43, 7, and 50%, respectively. In a control experiment, M.EcoGII-mediated A to 6mA conversion and M.CviPI-mediated GpC to Gp5mC conversion in phage λ DNA reached ∼93 and ∼30%, respectively (data not shown).

Base composition analysis of the *Eco*GII-treated phi W-14 genomic DNA indicated that 56% of adenosines were converted to 6mA (**Figure 4**). The C5-cytosine MTases M.CviPI and M.SssI were capable of converting 8% and 12% of cytidines to 5mC in phi W-14 DNA, respectively. The naturally occurring putT in phage phi W-14 DNA was detected at approximately 48%, which is consistent with the ∼50% putT reported in a previous work (Kropinski et al., 1973; Maltman et al., 1980) (note that total levels of putT reported here include the putT-G and putT-C dinucleotides, which result from the incomplete digestion due to the presence of the putrescinyl group) (**Figure 4**). The reason for poor methylation by the C5 MTases on ViI and phi W-14 DNA is unknown. Poor cytosine methylation may provide certain advantage against 5mC-dependent restriction systems such as *Bis*I, *Mcr*BC, *Mcr*A, *Msp*JI, and *Taq*I homologs (Cohen-Karni et al., 2011; Xu et al., 2016; Kisiala et al., 2018).

Ligation of Restriction Fragments of Phages ViI and phi W-14 DNA

In phage ViI DNA, approximately 43% of Ts have been replaced by 5-*NeO*mdU. The percentage of putT replacing T in phage phi W-14 was in the range of 47-48% (see **Figure 4**). We examined the ligation efficiency of restriction fragments from phage ViI and phi W-14 by T4 DNA ligase. *Nla*III- (CATG/) and *Fat*I- (/CATG) partially digested, or *Rsa*I (GT/AC) completely digested ViI restriction fragments were ligated at 16°C overnight. The sticky ends of *Nla*III and *Fat*I fragments were efficiently ligated, whereas the blunt-ended *Rsa*I fragments were ligated at a lower efficiency (**Figure 5A**). The *Sau*3AI- or *Mbo*I-digested (partial digestions) of phi W-14 restriction fragments were ligated efficiently indicated by the appearance of large concatenated DNA after ligation. Lower ligation efficiency was observed for blunt-ended *Rsa*I fragments (**Figure 5B**). We concluded that even though modified T could slow down restriction digestions by *Nla*III and *Fat*I for ViI genomic DNA, or by *Sau*3AI and *Mbo*I for phi W-14 DNA, the resulting restriction fragments can be efficiently ligated by T4 DNA ligase. The lower efficiency of *Rsa*I fragment ligation is most likely due to the blunt-ended nature of the ligation (Sambrook et al., 1989; Tsai et al., 2017).

Exonuclease Digestion of ViI and phi W-14 Genomic DNA

We next examined exonuclease activity on phage M6, ViI and phi W-14 DNA. Two types of phage DNA restriction fragments were digested with different amount of λ exonuclease or *E. coli* exonuclease III. Phage M6 and ViI restriction fragments were equally degraded by the two exonucleases (**Supplementary Figures 7A,B**). However, phi W-14 restriction fragments showed apparent slowed-down in exonuclease degradation (at 10–20 U range vs. 0.5 μg DNA) (**Supplementary Figure 7C**). Unmodified 2-log DNA ladder is sensitive to *E. coli* exonuclease III and λ exonuclease digestions (data not shown). The mechanism of phi W-14 DNA partial resistance to exonuclease digestion is still unknown. It was reported previously that the rate of

TABLE 3 | Methylation and subsequent restriction challenge of methylated phage M6 DNA.

Type of DNA MTase		Sequence modified	REase used to challenge DNA	Methylation status	Cleavage of gDNA prior to methylation
Phage M6	**5mC MTase**				
	M.AluI	AGCT	*Alu*I	Partial	Partial
	CpG (M.SssI)	CG	*Hpa*II	Partial	Complete
	GpC (M.CviPI)	GC	*Hha*I	Complete	Complete
	M.HaeIII	GGCC	*Hae*III	Partial	Complete
	M.HhaI	GCGC	*Hha*I	Partial	Complete
	M.HpaII	CCGG	*Hpa*II	Complete	Complete
	M.MspI	CCGG	*Msp*I	Complete	Complete
	6mA MTase				
	Dam	GATC	*Mbo*I	Partial	Complete
	M.EcoGII	A	*Mbo*I	Partial	Complete
	M.TaqI	TCGA	*Taq*I	Partial	Complete

Highlighted sequences, complete methylation.

DNA hydrolysis by non-specific endonuclease of modified phage PBS1 (dT substituted by dU) was decreased by 14.3-fold, and hypermodified phage T4 DNA also shows slow-down in nuclease degradation (Huang et al., 1982).

Digestion of Phage ViI and phi W-14 DNA With DNA Glycosylase and AP Endonuclease

5hmdU can be excised by DNA repair enzymes AlkA and Mug from *E. coli*, and by human SMUG1 (hSMUG1) and TDG to create AP sites (apurinic/apyrimidinic site), which can be further cleaved by AP endonucleases (Ulbert et al., 2004). Since ViI genomic DNA contains a small amount of 5hmdU we tested whether ViI and phi W-14 genomic DNA could be fragmented by hSMUG1 and AP endonuclease. **Supplementary Figure 8** shows that a small amount of smearing of ViI gDNA after treatment with hSMUG1 and Endonuclease VIII, probably resulting from cleavage in the small percentage of 5hmdU in the genome. Phi W-14 and λ DNA (a negative control) is quite resistant to the cleavage by the combination of these two enzymes. In the positive control sample, phage SP8 DNA (5hmdU substituted for T) was extensively hydrolyzed by hSMUG1 and Endonuclease VIII.

DISCUSSION

Biological Function of Base Modification (nt Substitution)

In bacterial host and phage coevolution, phage use extensive base modifications (nt substitutions) to protect its genome against host restrictions. The results presented here demonstrate that hypermodified T derived from 5hmdU can also efficiently protect phage genomes against Type II restrictions, in analogous manner to modified Gs, such as dG^{+} found in phage 9g genome (Thiaville et al., 2016) and 2′-deoxy-7-amido-7-deazaguanosine (dADG) found in certain bacteria genomic islands (Yuan et al., 2018), to modified As, such as *N*6-(1-acetamido)-adenine in phage Mu genome (Hattman, 1979), to modified Cs, such as 5gmC in phage T4, 5hmC in phage T4gt, and 5mC in phage XP12 genome. Although not much *in vivo* restriction study has been carried out on T-hypermodified phages, it is very likely that there is a strong correlation between *in vitro* and *in vivo* restriction activity. Depending on the *in vivo* enzyme activity and level of restriction gene expression, restriction of phage infection can be in the range of 10^{2}–10^{6} fold (reviewed in Pingoud et al., 2016). In this work we focus on Type II restrictions *in vitro*. Resistance against Type I restriction has not been studied and we only tested one ATP-dependent Type III restriction (*Eco*P15I, CAGCAG N25/). We hypothesize that phages M6, ViI, and phi W-14 may be resistant or partially resistant to Type I restriction as long as the restriction sites of these enzymes contain one or more TS dinucleotide sequence. 5mC-dependent REases are not tested on the three phage DNA substrates.

To counter adenine or cytosine modifications of phage genomes, bacteria develop modification-dependent REases (MDRE) to specifically attack modified DNA (Fleischman et al., 1976; Raleigh et al., 1989). For example, the *E. coli* GmrSD endonuclease attacks 5hmC and 5gmC modified DNA (Bair and Black, 2007; He et al., 2015). We have not found MDREs against modified T or modified G, but such enzymes might exist in nature. In addition, phages use anti-restriction proteins, small inhibitor proteins, DNA mimic protein to inhibit host-encoded REase (Rifat et al., 2008). Another likely function of modified bases is to help phage DNA packaging; for example, the positive charges of protonated $-NH_2$ groups in the putrescinyl group of putT side chain helps counter balance the negative charges of the DNA backbone, thereby enhancing DNA structural flexibility and denser packing the DNA into the viral capsid (Scraba et al., 1983). In addition to enhanced DNA packing capability, modified bases have also been implicated in regulation of promoter strength and gene expression during initiation of DNA packaging into phage prohead (Greene et al., 1986). This has been demonstrated in phage P1 that the GATC sequences in the packaging site (*pac*) are recognized and methylated by the phage-encoded Dam MTase triggering cleavage of *pac* sites and phage packaging initiation (Coulby and Sternberg, 1987;

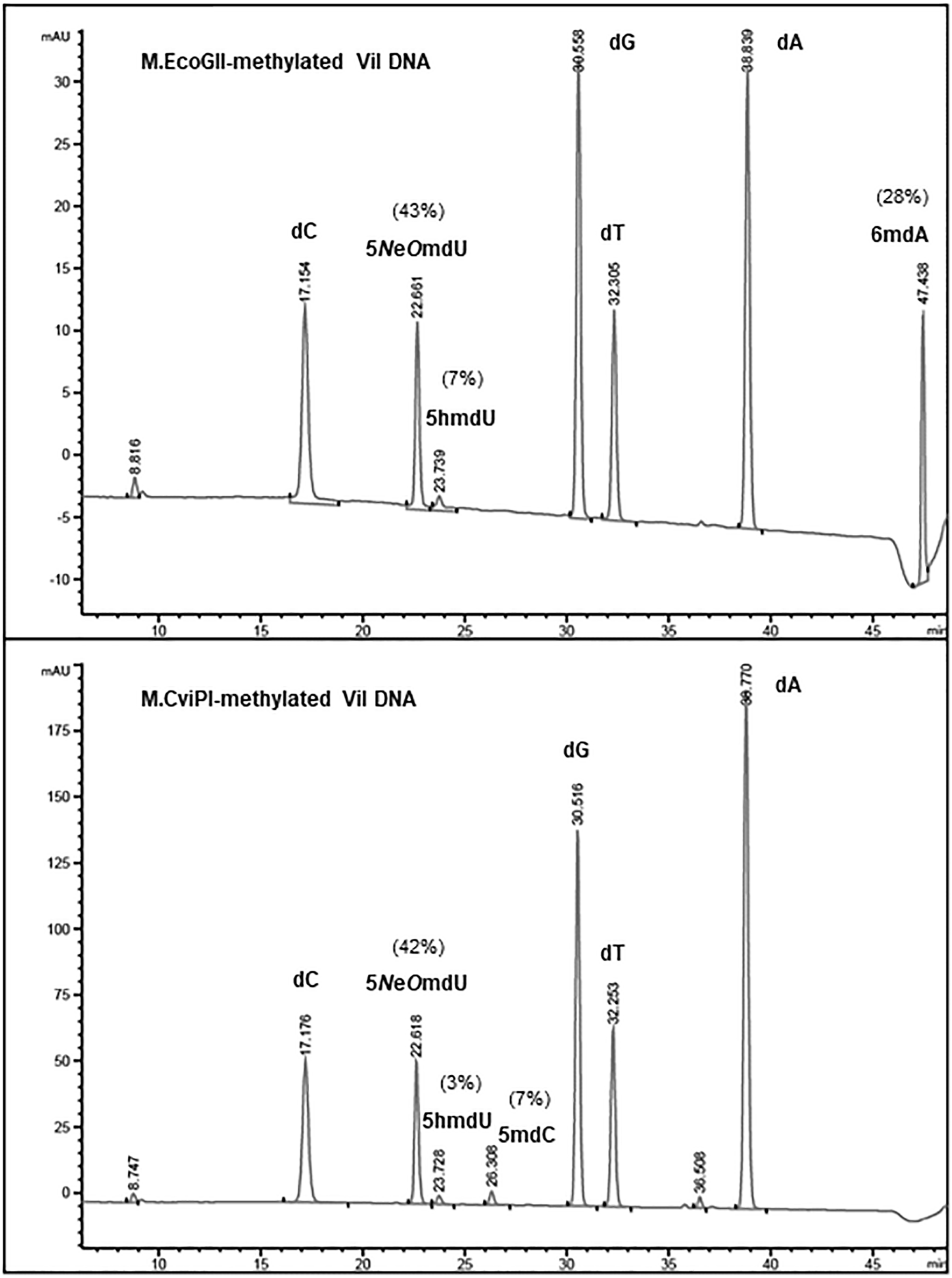


FIGURE 3 | Base composition analysis by LC-MS of M.EcoGII- or M.CviPI-methylated Vil gDNA. The distribution of thymine-derived bases 5-*NeO*mdU, 5hmdU, and dT were estimated at 43, 7, and 50% in phage Vil genome, respectively. The percentage of 6mA generated by treatment with M.EcoGII was approximately 28% of the total adenosines (**top** panel). The percentage of 5mC was only 7% after methylation of the phage Vil gDNA with the GpC methyltransferase M.CviPI (**bottom** panel).

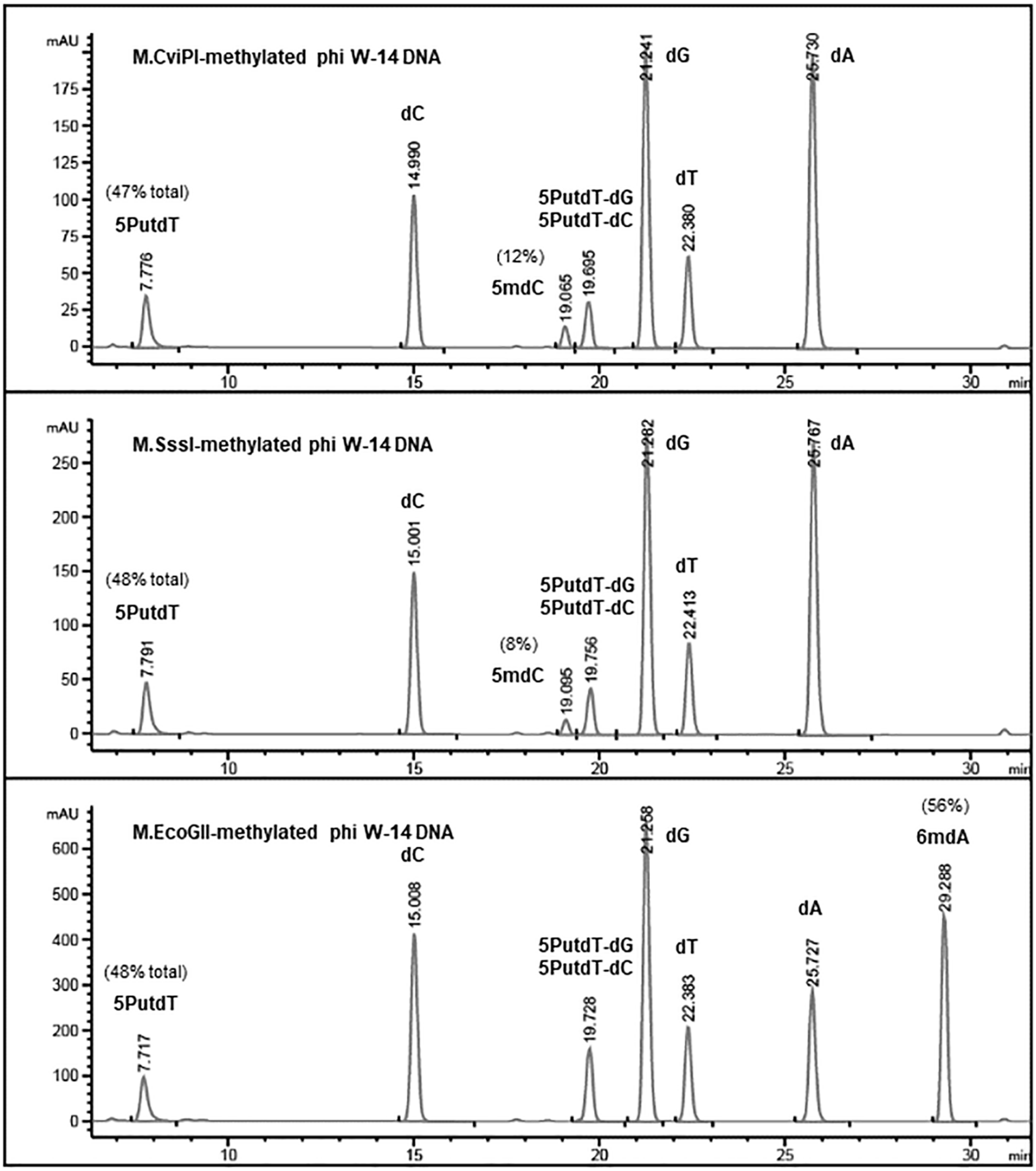


FIGURE 4 | Base composition analysis by LC-MS of M.CviPI-, M.SssI-, and M.EcoGII-methylated phi W-14 genomic DNA. The percentage of 5mC was ∼8 and ∼12% after treatment with the GpC and CpG methyltransferases M.CviPI and M.SssI, respectively. 6mA levels reached 56% after methylation with M.EcoGII. The natural modified base putT was detected in the range of 47–48%, in close agreement with previously published results (∼50%). A small fraction of putT was present in the form putT-G and putT-C dinucleotides due to incomplete digestion of the phi W-14 DNA.

Sternberg and Coulby, 1990). The effect of modified Ts on promoter strength and transcription regulation remains to be studied for the three phages reported here.

Bacillus phage SPO1 genomic DNA wherein >98% of Ts are replaced by 5hmdU is resistant or partially resistant to over 50% of Type II REases with 0–4 Ts in the recognition sequences (Huang et al., 1982; Vilpo and Vilpo, 1995). Another important aspect of the non-canonical nucleotide in the genome is the regulation of viral gene expression: temporal differential expression of the early and late viral genes in transcription (Greene et al., 1986; Hoet et al., 1992). Phage M6 DNA carrying the modified base 5-aminoethyl which confers slightly higher

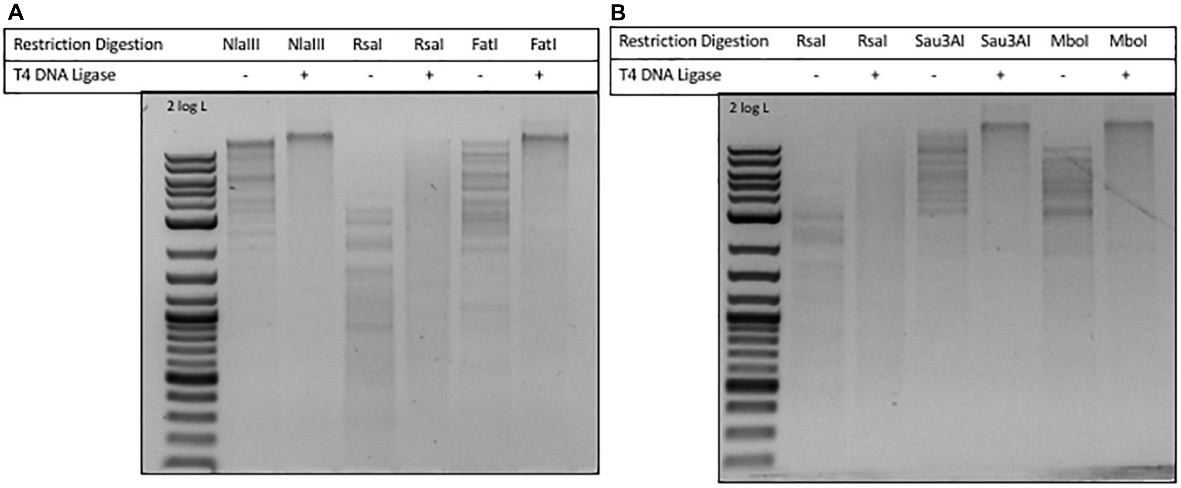


FIGURE 5 | Ligation of phages ViI and phi W-14 restriction fragments by T4 DNA ligase. **(A)** Ligation of phage ViI restriction fragments partially digested by *Nla*III and *Fat*I, and completely digested by *Rsa*I. Cleavage sites of REases: *Nla*III (CATG/), 3′ 4-base overhang; *Rsa*I (GT/AC), blunt ends; *Fat*I (/CATG), 5′ 4-base overhang. **(B)** Ligation of phage phi W-14 restriction fragments generated by *Rsa*I, *Sau*3AI, and *Mbo*I digestions. Cleavage sites of *Rsa*I (GT/AC), blunt ends; *Sau*3AI (/GATC), and *Mbo*I (/GATC), 5′ 4-base overhang.

resistance (48.9% complete and 15.8% partial resistance). More complex modifications, such as in phage ViI led to even higher resistance level (∼71.0%). It is not clear how phages balance the need for base modification to become highly resistant to host-encoded restrictions and energy (ATP) consumption on making these base modifications and the ultimate evolutionary advantage in successful infection of bacterial hosts. Some *Bacillus* phage or prophage genomes encode frequent multi-specificity cytosine MTases (Xu et al., 1992, 1997; Schumann et al., 1995). Phage T2 and T4 encode an adenine MTase (dam^+) that methylates GATC sites to provide more resistance against REases with overlapping GATC sequence. T even phages provide examples of two types of base modifications (6mA+5hmC or 6mA+5gmC) in their genome (Schlagman and Hattman, 1983). We have not yet observed phage genomes having both modified cytosine and thymine perhaps because of the small sample size of the sequenced phage genomes. Phage λ DNA contains some modified cytosine (5mC) and adenine (6mA, ∼15%) when the phage is propagated on Dam^+ Dcm^+ *E. coli* host. With advancement in DNA sequencing technology, single molecule SMRT sequencing and Nanopore sequencing might be able to sequence and identify more modifies bases in addition to N4mC and N6mA in DNA (Flusberg et al., 2010; Clark et al., 2012).

MDRE in *Pseudomonas* Strains and Phage Therapy

The *Pseudomonas* phage M6 DNA can be efficiently methylated by a few frequent C5-cytosine MTases to achieve double base modifications, which can provide more protection against Type II restrictions with GC recognition sequence. But the 5mC modifications also provide an opportunity for 5mC-dependent restrictions. Some *Pseudomonas* genomes encode *Mcr*BC and Mrr-like, and *Bis*I-like enzymes (REBASE) that remain to be characterized.

A cocktail of *Pseudomonas* lytic phages has been successfully used to treat *P. aeruginosa* infections in animal models (Forti et al., 2018). The DNA restriction data presented here suggests clinicians should take into consideration of heavily modified phage genomes and host restriction systems on the success or failure of phage-based therapies.

Conserved Sequence Motif in Resistant Sites

Analysis of the resistant sites in the phage genomes revealed a conserved motif TG, TC, or TS, suggesting the modified Ts possess certain sequence specificity, which may have been conferred by phage DNA 5hmdU kinase that phosphorylates the base for further chemical modification. Understanding the enzymes involved in thymidine hypermodification in phage genomes is an active research topic in our lab (YJL, PW) (Lee et al., 2018). In support of the preferred TS specificity observed among the resistant sites, purified DNA 5hmdUMP kinase can phosphorylate the 5hmdU base in phage DNA substrates (*Nco*I, CCA$\underline{\text{TG}}$G) (PW, unpublished result). For complete restriction digestion of hypermodified T phage DNA and cloning of certain genes (restriction fragments), **Supplementary Tables 1–3** provide a useful guidance to choose among various commercially available restriction enzymes.

CRISPR-Cas Associated Protein Cas4 Nuclease and Homing Endonucleases

Both ViI and phi W-14 encode a three-gene cluster with predicted function in restriction (phage against phage superinfection). ORFs Vi01_137, 138, and 139 encode putative RNA-DNA and DNA-DNA helicase/ATPase, CRISPR-Cas associated protein Cas4 nuclease (Cas4 IA-ID, IIB superfamily), and ssDNA binding protein in the ViI genome. Similarly, a three-gene cluster in phi W-14 genome contains gp030, gp032, and gp031. But the exact function of the three proteins involved in DNA metabolism (restriction) is still unknown. Incidentally, phage ViI also encodes a superinfection exclusion protein (Vi01_111c) that may play a role in attenuation of other phage infections. ViI genome encodes one GIY-YIG superfamily endonuclease (Vi01_159c); and phi W-14 genome encodes two HNH endonucleases (gp143 and gp219). These endonucleases are probably homing endonucleases involved in insertion of intron into intronless target known as intron "homing" since there are no cognate MTase genes associated with the predicted endonucleases (reviewed in Stoddard and Belfort, 2010). Because of large recognition sequence of homing endonucleases (typically 16–30 bp), there is no need to encode cognate MTase for self-protection.

Base J and 5hmdU in Eukaryotic Parasite, DNA Glycosylase/AP Endonuclease

Base J (*O*-linked glucosylated thymine, β-D-glucosyl -deoxymethyluracil) in human pathogens *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania* species, consisted of about 1% of total T in the genomes. The modified base J is an important regulatory epigenetic mark in trypanosomatids to influence gene expression. The JBP1/2 enzymes catalyze hydroxylation of thymine (Yu et al., 2007), forming 5-hydroxymethyluracil (5hmdU), which is then glucosylated by the base J-associated glucosyltransferase (JGT). The presence of glucosylated 5hmdU has not been reported in phage genomes. DNA with base J modification is not a substrate for DNA repair enzyme AlkA and Mug of *E. coli*, and hSMUG1 and TDG (Ulbert et al., 2004). When phage genomes contain large number of 5hmdU bases, the phage DNA is possibly subjected to DNA glycosylase cleavage. We show that phage ViI gDNA can be partially digested by hSMUG1 and Endonuclease VIII due to the presence of small amount of 5hmdU in the genome. But phi W-14 is fully resistant to hSMUG1 and endonuclease VIII; while phage SP8 genome with 5hmdU is heavily degraded by the two enzymes. The 5hmdU base is to be further modified to become resistant to host DNA glycosylases/AP endonucleases and REases such as in the case of phage M6, ViI and phi W-14. Alternatively, the 5hmdU-containing phages can only infect bacterial hosts deficient in AlkA- and Mug-like repair enzymes or by expression of phage-encoded enzyme inhibitors.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

KF, RT, MX, IC, AC, and S-YX performed experimental work. M-QX, Y-JL, IC, PW, and S-YX contributed with ideas. S-YX wrote the manuscript with input from all the authors.

FUNDING

KF, RT, MX, and AC were supported by NEB's summer research and internship program. The financial support for this work and open access fee was provided by New England Biolabs, Inc.

ACKNOWLEDGMENTS

We thank Bill Jack and Rich Roberts for their critical comments on the manuscript, and Andy Gardner and Lise Raleigh for helpful discussions. We also thank Don Comb, Jim Ellard, Rich Roberts, Andy Gardner, Bill Jack, and Tom Evans for their guidance and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.00584/full#supplementary-material

REFERENCES

Bair, C. L., and Black, L. W. (2007). A type IV modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified DNAs. *J. Mol. Biol.* 366, 768–778. doi: 10.1016/j.jmb.2006.11.051

Bertani, G., and Weigle, J. J. (1953). Host controlled variation in bacterial viruses. *J. Bacteriol.* 65, 113–121.

Carson, S., Wilson, J., Aksimentiev, A., Weigele, P. R., and Wanunu, M. (2016). Hydroxymethyluracil modifications enhance the flexibility and hydrophilicity of double-stranded DNA. *Nucleic Acids Res.* 44, 2085–2092. doi: 10.1093/nar/gkv1199

Clark, T. A., Murray, I. A., Morgan, R. D., Kislyuk, A. O., Spittle, K. E., Boitano, M., et al. (2012). Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *Nucleic Acids Res.* 40, e29. doi: 10.1093/nar/gkr1146

Cohen-Karni, D., Xu, D., Apone, L., Fomenkov, A., Sun, Z., Davis, P. J., et al. (2011). The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11040–11045. doi: 10.1073/pnas.1018448108

Coulby, J., and Sternberg, N. (1987). Bacteriophage P1 encodes its own dam methylase. *Plasmid* 17, 81.

Czapinska, H., Kowalska, M., Zagorskaite, E., Manakova, E., Slyvka, A., and Xu, S. Y. (2018). Activity and structure of EcoKMcrA. *Nucleic Acids Res.* 46, 9829–9841. doi: 10.1093/nar/gky731

Dila, D., Sutherland, E., Moran, L., Slatko, B., and Raleigh, E. A. (1990). Genetic and sequence organization of the mcrBC locus of *Escherichia coli* K-12. *J. Bacteriol.* 172, 4888–4900. doi: 10.1128/jb.172.9.4888-4900.1990

Drozdz, M., Piekarowicz, A., Bujnicki, J. M., and Radlinska, M. (2012). Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res.* 40, 2119–2130. doi: 10.1093/nar/gkr1039

Feng, T. Y., Tu, J., and Kuo, T. T. (1978). Characterization of deoxycytidylate methyltransferase in Xanthomonas oryzae infected with bacteriophage Xp12. *Eur. J. Biochem.* 87, 29–36. doi: 10.1111/j.1432-1033.1978.tb12348.x

Fleischman, R. A., Cambell, J. L., and Richardson, C. C. (1976). Modification and restriction of T-even bacteriophages. In vitro degradation of deoxyribonucleic acid containing 5-hydroxymethylctosine. *J. Biol. Chem.* 251, 1561–1570.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465. doi: 10.1038/nmeth.1459

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:e2573-17. doi: 10.1128/AAC.02573-17

Gold, L. M., and Schweiger, M. (1969). Synthesis of phage-specific alpha- and beta-glucosyl transferases directed by T-even DNA in vitro. *Proc. Natl. Acad. Sci. U.S.A.* 62, 892–898. doi: 10.1073/pnas.62.3.892

Greene, J. R., Morrissey, L. M., and Geiduschek, E. P. (1986). DNA binding by the bacteriophage SPO1-encoded type II DNA-binding protein, transcription factor 1. Site-specific binding requires 5-hydroxymethyluracil-containing DNA. *J. Biol. Chem.* 261, 12828–12833.

Hattman, S. (1979). Unusual modification of bacteriophage Mu DNA. *J. Virol.* 32, 468–475.

He, X., Hull, V., Thomas, J. A., Fu, X., Gidwani, S., Gupta, Y. K., et al. (2015). Expression and purification of a single-chain Type IV restriction enzyme Eco94GmrSD and determination of its substrate preference. *Sci. Rep.* 5: 9747. doi: 10.1038/srep09747

Heitman, J., and Model, P. (1987). Site-specific methylases induce the SOS DNA repair response in *Escherichia coli*. *J. Bacteriol.* 169, 3243–3250. doi: 10.1128/jb.169.7.3243-3250.1987

Hoet, P. P., Coene, M. M., and Cocito, C. G. (1992). Replication cycle of *Bacillus subtilis* hydroxymethyluracil-containing phages. *Annu. Rev. Microbiol.* 46, 95–116. doi: 10.1146/annurev.mi.46.100192.000523

Hoskisson, P. A., Sumby, P., and Smith, M. C. (2015). The phage growth limitation system in Streptomyces coelicolor A(3)2 is a toxin/antitoxin system, comprising enzymes with DNA methyltransferase, protein kinase and ATPase activity. *Virology* 477, 100–109. doi: 10.1016/j.virol.2014.12.036

Huang, L. H., Farnet, C. M., Ehrlich, K. C., and Ehrlich, M. (1982). Digestion of highly modified bacteriophage DNA by restriction endonucleases. *Nucleic Acids Res.* 10, 1579–1591. doi: 10.1093/nar/10.5.1579

Iyer, L. M., Zhang, D., Burroughs, A. M., and Aravind, L. (2013). Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. *Nucleic Acids Res.* 41, 7635–7655. doi: 10.1093/nar/gkt573

Kelln, R. A., and Warren, R. A. (1973). Studies on the biosynthesis of alpha-putrescinylthymine in bacteriophage phi W-14-infected *Pseudomonas acidovorans*. *J. Virol.* 12, 1427–1433.

Kisiala, M., Copelas, A., Czapinska, H., Xu, S. Y., and Bochtler, M. (2018). Crystal structure of the modification-dependent SRA-HNH endonuclease TagI. *Nucleic Acids Res.* 46, 10489–10503. doi: 10.1093/nar/gky781

Kropinski, A. M., Bose, R. J., and Warren, R. A. (1973). 5-(4-Aminobutylaminomethyl)uracil, an unusual pyrimidine from the deoxyribonucleic acid of bacteriophage phiW-14. *Biochemistry* 12, 151–157. doi: 10.1021/bi00725a025

Kruger, D. H., and Bickle, T. A. (1983). Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiol. Rev.* 47, 345–360.

Kulikov, E. E., Golomidova, A. K., Letarova, M. A., Kostryukova, E. S., Zelenin, A. S., Prokhorov, N. S., et al. (2014). Genomic sequencing and biological characteristics of a novel *Escherichia coli* bacteriophage 9g, a putative representative of a new Siphoviridae genus. *Viruses* 6, 5077–5092. doi: 10.3390/v6125077

Landy, A., Ruedisueli, E., Robinson, L., Foeller, C., and Ross, W. (1974). Digestion of deoxyribonucleic acids from bacteriophage T7, lambda, and Phi80h with site-specific nucleases from Hemophilus influenzae strain Rc and strain Rd. *Biochemistry* 13, 2134–2142. doi: 10.1021/bi00707a022

Lee, Y. J., Dai, N., Walsh, S. E., Muller, S., Fraser, M. E., and Kauffman, K. M. (2018). Identification and biosynthesis of thymidine hypermodifications in the genomic DNA of widespread bacterial viruses. *Proc. Natl. Acad. Sci. U.S.A.* 115, E3116–E3125. doi: 10.1073/pnas.1714812115

Luria, S. E. (1953). Host-induced modifications of viruses. *Cold Spring Harbor Symp. Quant. Biol.* 18, 237–244. doi: 10.1101/SQB.1953.018.01.034

Luria, S. E., and Human, M. L. (1952). A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol.* 64, 557–569.

Maltman, K. L., Neuhard, J., Lewis, H. A., and Warren, R. A. (1980). Synthesis of thymine and alpha-putrescinylthymine in bacteriophage phi W-14-infected *Pseudomonas* acidovorans. *J. Virol.* 34, 354–359.

Matsuoka, S., Asai, K., and Sadaie, Y. (2005). Restriction and modification of SP10 phage by BsuM of Bacillus subtilis Marburg. *FEMS Microbiol. Lett.* 244, 335–339. doi: 10.1016/j.femsle.2005.02.006

Miller, P. B., Wakarchuk, W. W., and Warren, R. A. (1985). Alpha-putrescinylthymine and the sensitivity of bacteriophage Phi W-14 DNA to restriction endonucleases. *Nucleic Acids Res.* 13, 2559–2568. doi: 10.1093/nar/13.7.2559

Mulligan, E. A., and Dunn, J. J. (2008). Cloning, purification and initial characterization of *E. coli* McrA, a putative 5-methylcytosine-specific nuclease. *Protein Exp. Purif.* 62, 98–103. doi: 10.1016/j.pep.2008.06.016

Murray, I. A., Morgan, R. D., Luyten, Y., Fomenkov, A., Corrêa, IR Jr, Dai, N., et al. (2018). The non-specific adenine DNA methyltransferase M.EcoGII. *Nucleic Acids Res.* 46, 840–848. doi: 10.1093/nar/gkx1191

Neuhard, J., Maltman, K. L., and Warren, R. A. (1980). Bacteriophage phi W-14-infected *Pseudomonas acidovorans* synthesizes hydroxymethyldeoxyuridine triphosphate. *J. Virol.* 34, 347–353.

Pingoud, A., Wilson, G. G., and Wende, W. (2016). Type II restriction endonucleases - a historical perspective and more. *Nucleic Acids Res.* 44:8011. doi: 10.1093/nar/gkw513

Raleigh, E. A., Trimarchi, R., and Revel, H. (1989). Genetic and physical mapping of the mcrA (rglA) and mcrB (rglB) loci of *Escherichia coli* K-12. *Genetics* 122, 279–296.

Rifat, D., Wright, N. T., Varney, K. M., Weber, D. J., and Black, L. W. (2008). Restriction endonuclease inhibitor IPI* of bacteriophage T4: a novel structure for a dedicated target. *J. Mol. Biol.* 375, 720–734. doi: 10.1016/j.jmb.2007.10.064

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., et al. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 31, 1805–1812. doi: 10.1093/nar/gkg274

Robinson, C. R., and Sligar, S. G. (1993). Molecular recognition mediated by bound water. A mechanism for star activity of the restriction endonuclease EcoRI. *J. Mol. Biol.* 234, 302–306. doi: 10.1006/jmbi.1993.1586

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning, A Laboratory Manual.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Sawaya, M. R., Zhu, Z., Mersha, F., Chan, S. H., Dabur, R., Xu, S. Y., et al. (2013). Crystal structure of the restriction-modification system control element C.Bcll and mapping of its binding site. *Structure* 13, 1837–1847. doi: 10.1016/j.str.2005.08.017

Schlagman, S. L., and Hattman, S. (1983). Molecular cloning of a functional dam+ gene coding for phage T4 DNA adenine methylase. *Gene* 22, 139–156. doi: 10.1016/0378-1119(83)90098-7

Schumann, J., Willert, J., Wild, C., Waler, J., and Trautner, T. A. (1995). M.B*ssHII: a new multispecific C*5-DNA-methyltransferase. *Gene* 157, 103–104. doi: 10.1016/0378-1119(95)00723-J

Scraba, D. G., Bradley, R. D., Leyritz-Wills, M., and Warren, R. A. (1983). Bacteriophage phi W-14: the contribution of covalently bound putrescine to DNA packing in the phage head. *Virology* 124, 152–160. doi: 10.1016/0042-6822(83)90298-2

Smith, H. O., and Wilcox, K. W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J. Mol. Biol.* 51, 379–391. doi: 10.1016/0022-2836(70)90149-X

Sternberg, N., and Coulby, J. (1990). Cleavage of the bacteriophage P1 packaging site (pac) is regulated by adenine methylation. *Proc. Natl. Acad. Sci. U.S.A.* 87, 8070–8074. doi: 10.1073/pnas.87.20.8070

Stewart, C. R., Casjens, S. R., Cresawn, S. G., Houtz, J. M., Smith, A. L., Ford, M. E., et al. (2009). The genome of Bacillus subtilis bacteriophage SPO1. *J. Mol. Biol.* 388, 48–70. doi: 10.1016/j.jmb.2009.03.009

Stoddard, B., and Belfort, M. (2010). Social networking between mobile introns and their host genes. *Mol. Microbiol.* 78, 1–4. doi: 10.1111/j.1365-2958.2010.07217.x

Tao, T., and Blumenthal, R. M. (1992). Sequence and characterization of pvuIIR, the PvuII endonuclease gene, and of pvuIIC, its regulatory gene. *J. Bacteriol.* 174, 3395–3398. doi: 10.1128/jb.174.10.3395-3398.1992

Thiaville, J. J., Kellner, S. M., Yuan, Y., Hutinet, G., Thiaville, P. C., Jumpathong, W., et al. (2016). Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1452–E1459. doi: 10.1073/pnas.1518570113

Tsai, R., Correa, I. R., Xu, M. Y., and Xu, S. Y. (2017). Restriction and modification of deoxyarchaeosine (dG+)-containing phage 9 g DNA. *Sci. Rep.* 7:8348. doi: 10.1038/s41598-017-08864-4

Ulbert, S., Eide, L., Seeberg, E., Borst, P., and Base, J. (2004). found in nuclear DNA of Trypanosoma brucei, is not a target for DNA glycosylases. *DNA Repair.* 3, 145–154. doi: 10.1016/j.dnarep.2003.10.009

Vasu, K., Nagamalleswari, E., Zahran, M., Imhof, P., Xu, S. Y., Zhu, Z., et al. (2013). Increasing cleavage specificity and activity of restriction endonuclease KpnI. *Nucleic Acids Res.* 41, 9812–9824. doi: 10.1093/nar/gkt734

Vilpo, J. A., and Vilpo, L. M. (1995). Restriction, methylation and ligation of 5-hydroxymethyluracil-containing DNA. *Mutat. Res.* 316, 123–131. doi: 10.1016/0921-8734(95)90005-5

Vincze, T., Posfai, J., and Roberts, R. J. (2003). NEBcutter: a program to cleave DNA with restriction enzymes. *Nucleic Acids Res.* 31, 3688–3691. doi: 10.1093/nar/gkg526

Weigele, P., and Raleigh, E. A. (2016). Biosynthesis and function of modified bases in bacteria and their viruses. *Chem. Rev.* 116, 12655–12687. doi: 10.1021/acs.chemrev.6b00114

Xu, S. Y., Klein, P., Degtyarev, S., and Roberts, R. J. (2016). Expression and purification of the modification-dependent restriction enzyme BisI and its homologous enzymes. *Sci. Rep.* 6:28579. doi: 10.1038/srep28579

Xu, S. Y., Nugent, R. L., Kasamkattil, J., Fomenkov, A., Gupta, Y., Aggarwal, A., et al. (1992). Characterization of type II and III restriction-modification systems from Bacillus cereus strains ATCC 10987 and ATCC 14579. *J. Bacteriol.* 194, 49–60. doi: 10.1128/JB.06248-11

Xu, S.-Y., Xiao, J.-P., Posfai, J., Maunus, R., and Benner, J. (1997). Cloning of the BssHII restriction-modification system in *Escherichia coli*: BssHII methyltransferase contains circularly permuted cytosine-5 methyltransferase motifs. *Nucleic Acids Res.* 25, 3991–3994. doi: 10.1093/nar/25.20.3991

Yu, Z., Genest, P. A., ter Riet, B., Sweeney, K., DiPaolo, C., Kieft, R., et al. (2007). The protein that binds to DNA base J in trypanosomatids has features of a thymidine hydroxylase. *Nucleic Acids Res.* 35, 2107–2115. doi: 10.1093/nar/gkm049

Yuan, Y., Hutinet, G., Valera, J. G., Hu, J., Hillebrand, R., Gustafson, A., et al. (2018). Identification of the minimal bacterial 2'-deoxy-7-amido-7-deazaguanine synthesis machinery. *Mol. Microbiol.* 110, 469–483. doi: 10.1111/mmi.14113

Conflict of Interest Statement: Y-JL, M-QX, IC, PW, and S-YX are employees of New England Biolabs, Inc. New England Biolabs commercializes reagents for molecular biological applications.

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.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 07 May 2019
doi: 10.3389/fmicb.2019.01001

Genome Analysis of Coxsackievirus A4 Isolates From Hand, Foot, and Mouth Disease Cases in Shandong, China

Min Wang[1†], Juan Li[1†], Ming-Xiao Yao[2†], Ya-Wei Zhang[3†], Tao Hu[1], Michael J. Carr[4,5], Sebastián Duchêne[6], Xing-Cheng Zhang[1], Zhen-Jie Zhang[1], Hong Zhou[1], Yi-Gang Tong[3], Shu-Jun Ding[2], Xian-Jun Wang[2]* and Wei-Feng Shi[1]*

[1] Key Laboratory of Etiology and Epidemiology of Emerging Infectious Diseases in Universities of Shandong, Taishan Medical College, Tai'an, China, [2] Shandong Provincial Key Laboratory of Communicable Disease Control and Prevention, Institute for Viral Disease Control and Prevention, Shandong Center for Disease Control and Prevention, Jinan, China, [3] State Key Laboratory of Pathogen and Biosecurity, Beijing Institute of Microbiology and Epidemiology, Beijing, China, [4] Global Station for Zoonosis Control, Global Institution for Collaborative Research and Education, Hokkaido University, Sapporo, Japan, [5] National Virus Reference Laboratory, School of Medicine, University College Dublin, Dublin, Ireland, [6] Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Parkville, VIC, Australia

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana–Champaign, United States

Reviewed by:
Koo Nagasawa,
Eastern Chiba Medical Center, Japan
Daniel C. Pevear,
VenatoRx Pharmaceuticals,
United States

***Correspondence:**
Xian-Jun Wang
xjwang62@163.com
Wei-Feng Shi
shiwf@ioz.ac.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: 31 January 2019
Accepted: 18 April 2019
Published: 07 May 2019

Citation:
Wang M, Li J, Yao M-X, Zhang Y-W, Hu T, Carr MJ, Duchêne S, Zhang X-C, Zhang Z-J, Zhou H, Tong Y-G, Ding S-J, Wang X-J and Shi W-F (2019) Genome Analysis of Coxsackievirus A4 Isolates From Hand, Foot, and Mouth Disease Cases in Shandong, China. Front. Microbiol. 10:1001. doi: 10.3389/fmicb.2019.01001

Coxsackievirus A4 (CVA4) is one of the most prevalent pathogens associated with hand, foot and mouth disease (HFMD), an acute febrile illness in children, and is also associated with acute localized exanthema, myocarditis, hepatitis and pancreatitis. Despite this, limited CVA4 genome sequences are currently available. Herein, complete genome sequences from CVA4 strains (n = 21), isolated from patients with HFMD in Shandong province, China between 2014 and 2016, were determined and phylogenetically characterized. Phylogenetic analysis of the *VP1* gene from a larger CVA4 collection (n = 175) showed that CVA4 has evolved into four separable genotypes: A, B, C, and D; and genotype D could be further classified in to two sub-genotypes: D1 and D2. Each of the 21 newly described genomes derived from isolates that segregated with sub-genotype D2. The CVA4 genomes displayed significant intra-genotypic genetic diversity with frequent synonymous substitutions occurring at the third codon positions, particularly within the P2 region. However, *VP1* was relatively stable and therefore represents a potential target for molecular diagnostics assays and also for the rational design of vaccine epitopes. The substitution rate of *VP1* was estimated to be 5.12×10^{-3} substitutions/site/year, indicative of ongoing CVA4 evolution. Mutations at amino acid residue 169 in *VP1* gene may be responsible for differing virulence of CVA4 strains. Bayesian skyline plot analysis showed that the population size of CVA4 has experienced several dynamic fluctuations since 1948. In summary, we describe the phylogenetic and molecular characterization of 21 complete genomes from CVA4 isolates which greatly enriches the known genomic diversity of CVA4 and underscores the need for further surveillance of CVA4 in China.

Keywords: Coxsackievirus A4, hand, foot, and mouth disease, genotypes, phylogenetic analysis, *VP1*

INTRODUCTION

Human enteroviruses (EVs), non-enveloped, single-stranded RNA viruses, are taxonomically classified in the genus *Enterovirus* (Dalldorf, 1953), family *Picornaviridae* and consist of four species: EV-A, EV-B, EV-C, and EV-D (Knowles et al., 2012). Coxsackievirus A4 (CVA4), a member of the EV-A species, is currently composed of 25 types including the common pathogens enterovirus A71 (EVA71) and Coxsackievirus A16 (CVA16) which are the most prevalent agents isolated from pediatric cases of hand-foot-mouth disease (HFMD), a common contagious disease among children which sporadically occur worldwide (Wang et al., 2017).

The CVA4 prototype strain, High Point (GenBank accession number: AY421762), was first isolated from urban sewage during a poliomyelitis outbreak in North Carolina, United States in 1948 (Melnick, 1950; Oberste et al., 2004). In Carey and Myers (1969), another CVA4 strain was isolated from the serum of a child with fatal illness marked by severe central nervous system manifestations. In Estrada Gonzalez and Mas (1977), found that CVA4 was associated with the occurrence of acute polyradiculoneuritis. Notably, CVA4 was also reported as pathogens in HFMD outbreaks (Hu et al., 2011). Associations between CVA4 infection and herpangina (Cai et al., 2015), mucocutaneous lymph node syndrome (Ueda et al., 2015), and bilateral idiopathic retinal vasculitis (Mine et al., 2017) have also been reported. These clinical findings highlight that CVA4 infections exert a disease burden to global public health, especially for neonates and young children which necessitates a greater understanding of the genetic diversity of this pathogen.

Few complete genomes from CVA4 strains were sequenced prior to 2004 and were predominantly from Asia and the United States (Oberste et al., 2004). However, in 2004 and 2006, CVA4 caused two HFMD epidemics in Taiwan (Chu et al., 2011). After 2006, a greater number of CVA4 strains were isolated in Europe and Asia, with the majority from China. For example, seven samples from acute flaccid paralysis (AFP) patients in Shaanxi province collected between 2006 and 2010 were diagnosed with CVA4 infection (Wang et al., 2016). Since 2008, predominantly partial VP1 and a far smaller number of complete genome sequences of multiple CVA4 strains have been reported from clinical specimens from pediatric HFMD cases in mainland China, such as Gansu (Liu et al., 2009), Shenzhen (Hu et al., 2011; Cai et al., 2015), Guangdong (Zhen et al., 2014), and Beijing (Li et al., 2015).

The CVA4 RNA genome (~7434 nt) can be sub-divided into the 5′-untranslated region (UTR), 3′-UTR, and the open reading frame (ORF) (~6606 nt) encoding one polyprotein comprising four structural proteins (*VP4*, *VP2*, *VP3*, and *VP1* in region P1) and seven non-structural proteins (*2A*, *2B*, and *2C* in region P2, and *3A*, *3B*, *3C*, and *3D* in region P3). The *VP1* gene is typically employed for taxonomic classification of EV types within the genus *Enterovirus* (Palacios et al., 2002). Currently, few full-length genome sequences of CVA4 (n = 10) are available in the GenBank database, which limits the understanding of CVA4 genome evolution and necessitates a greater understanding of the genetic diversity in this pathogen to provide an evidence base for the rational development of molecular diagnostics, drug development and vaccine design.

In the present study, we describe the full-length genome sequences (n = 21) from CVA4 isolates from pediatric HFMD patients from Shandong province, China identified between 2014 and 2016 and the phylogenetic and molecular characterization of CVA4 *VP1* sequences (n = 175). These findings deepen our understanding of CVA4 diversity and evolution and have important implications to mitigate global disease burden attributable to this pathogen.

MATERIALS AND METHODS

Clinical Samples

In this study, stool specimens from children <6 years of age presenting with HFMD were collected between 2014 and 2016, and EV-positive clinical samples which were laboratory-confirmed as non-EVA71 and non-CVA16 infections were kindly provided by the Shandong Center for Disease Control and Prevention.

CVA4 Isolation and RNA Extraction From Clinical Samples

Human rhabdomyosarcoma (RD) cells were grown in MEM (minimal essential medium, Gibco) supplemented with 10% FBS (fetal bovine serum, Gibco), 50 IU/mL of penicillin and 50 μg/mL of streptomycin. The samples were propagated three times in human RD cells at 37°C in a humid atmosphere under 5% CO_2.

Total RNA was extracted from the RD supernatant after development of cytopathic effect (CPE) using the RNAiso Plus kit (TaKaRa, 9109). RNA was reverse transcribed into cDNA using random hexamers with the PrimeScript RT Reagent kit (TaKaRa, RR037A). TaqMan-based real-time PCR assays were performed to detect the presence of EV using an ABI 7500 Real-Time PCR System, as previously described (Zhang et al., 2017a,b). The EV positive samples (n = 21) were sequenced using the MiSeq high-throughput sequencing platform, and the sequencing data were analyzed using the CLC program to obtain the complete viral genome sequences. Non-EVA71 and non-CVA16 samples from HFMD cases were further tested by a CVA4-specific real-time PCR assay employing oligonucleotide primers and probe designed based on the *VP1* gene of the prototype CVA4 strain (GenBank accession number: AY421762): CVA4-F: TATGGGCTTTGTCCAACTCC, CVA4-R: GTCTAGGGACCCATGCCCTCACT, CVA4-probe: FAM-TGGGGACATTTTCAGCTAGAGTTGTGAGCAAG-BHQ1.

Phylogenetic Analysis of CVA4

For phylogenetic analysis, two datasets including all available complete genomic sequences (n = 10) and the full-length sequences of *VP1* gene (n = 154) of CVA4 strains were downloaded from GenBank. Multiple sequence alignments for the two datasets were performed using Muscle (Edgar, 2004), together with the Shandong CVA4 sequenced genomes

(n = 21). The nucleotide substitution model was chosen using jModeltest (Darriba et al., 2012), and the general time reversible (GTR) model was always the best model for analysis of both datasets. Phylogenetic analysis of the two datasets was conducted using RAxML v8.1.6 (Stamatakis et al., 2005), with the GTRGAMMA nucleotide substitution model with 1000 bootstrap replicates. The nucleotide and amino acid distances were estimated using MEGA 5 (Tamura et al., 2011).

Genetic Diversity Along the CVA4 Polyprotein Genes

The average pairwise genetic diversity along the CVA4 genomes was calculated using Phylip with a sliding window of 300 nucleotides and a step size of 50 nucleotides (Faria et al., 2017). To identify at which site of the codons that nucleotide variations mostly occur, the *Shannon entropy (Sn)* at positions 1, 2, and 3 of each codon of the aligned polyprotein genes and the *VP4*, *VP2*, *VP3*, *VP1*, *2A*, *2B*, *2C*, *3A*, *3B*, *3C*, and *3D* genes, respectively, were calculated as reported previously (Ramirez et al., 2013). The data processing was performed on the open source R environment (R Development Core Team, 2012), using the ggplot2 package.

Evolutionary Dynamics of Global CVA4 Strains

To determine the molecular clock-like structure in the CVA4 *VP1* data, the maximum likelihood tree of the *VP1* gene sequences estimated in the previous step was used to study the association between sequence divergence and sampling times using TempEst (Rambaut et al., 2016). To better understand the evolutionary dynamics of CVA4, the *VP1* sequence dataset (n = 175) was used to estimate the nucleotide substitution rate using Bayesian Markov chain Monte Carlo (MCMC) sampling implemented in the BEAST v1.8.4 package (Drummond and Rambaut, 2007). The SRD09 model was employed as the nucleotide substitution model. The Bayesian MCMC process was run for 100 million steps, with a sampling frequency of every 5,000 steps and the first 10% of the steps were discarded as burn-in. To select the best combination of the molecular clock and tree prior, both path sampling and stepping-stone sampling (Baele et al., 2012) were employed. The best model combination was a Bayesian skyline tree prior and an uncorrelated relaxed molecular clock model, with log-normally distributed variation in rates among branches (**Supplementary Table S1**) (Drummond et al., 2005). The posterior distributions were evaluated using Tracer v1.6[1] where the estimated sample size (ESS) was no less than 200. The Bayesian maximum clade credibility (MCC) tree was generated using TreeAnnotator v1.8.4, and then visualized using FigTree v1.4.3[2]. In addition, we performed a Bayesian skyline plot analysis using Tracer v1.6 program to reconstruct the evolutionary history based on the 175 full-length *VP1* gene sequences of CVA4 strains.

RESULTS

Dataset Summary

In this study, a total of 21 CVA4 strains were isolated from stool specimens of HFMD patients in Shandong Province, including five from Liaocheng, three from Heze, two each from Dezhou, Linyi, Weihai, Yantai and single samples from Binzhou, Laiwu, Taian, Qingdao and Rizhao (**Figure 1** and **Supplementary Table S2**). The whole genome sequences of the 21 isolates have been deposited into the GenBank database: MH086030-MH086050 (**Supplementary Table S2**).

Ten complete CVA4 genomes between 1948 and 2015 were publicly available from GenBank, with eight strains from China and two from the United States. Full-length CVA4 *VP1* gene sequences (n = 175) strains were also retrieved. They were isolated from the Republic of Azerbaijan (n = 1), China (n = 145), India (n = 8), Kenya (n = 1), Russia (n = 15), Turkmenistan (n = 1), and the United States (n = 4) (**Supplementary Figure S1**), respectively. The collection dates ranged from 1948 to 2016 (**Supplementary Figure S1**). Notably, 94.9% (n = 166) of them were sampled since 2006 and 84.9% (n = 141) were collected from 16 provinces and cities in China, including Shandong (n = 60), Sichuan (n = 18), Beijing (n = 14), Shanghai (n = 11), Shenzhen (n = 9), Hunan (n = 7), Yunnan (n = 4), Chongqing (n = 4), Jilin (n = 3), Ningxia (n = 2), Jiangsu (n = 2), Hainan (n = 2), Guangdong (n = 2), Anhui (n = 1), Henan (n = 1), and Zhejiang (n = 1).

Phylogenetic Characterization of CVA4

Phylogenetic analysis of the 31 whole genomes revealed that the 21 isolates from the present study segregated into three strongly supported clades (**Figure 1**). 17 strains were located in a major clade together with seven Chinese strains collected between 2008 and 2011. Two isolates, SDLC/16114/China/2016-05-08 and SDQD/QD384R/China/2016-06-17 clustered into one independent clade and the isolates SDLY/2014LY124R/China/2014-04-23 and SDHZ/H337/China/2015-04-28, as well as the previously described KP289442/P1033/2013/China/2013, belonged to a separable third clade. In addition, the 21 Shandong isolates shared higher nucleotide sequence homologies with eight previously described Chinese strains (82.0–99.8%) than with the prototype strain High Point/United States/1948 (77.8–79.3%) and the American strain, 2015-OB2038/TN/United States/2015-04-16 (75.5–78.3%).

In the maximum likelihood tree of *VP1* genes (**Supplementary Figure S2**), the CVA4 isolates were divided into four highly supported genotypes: A (n = 1), B (n = 1), C (n = 28), and D (n = 145). The prototype High Point strain isolated from the United States in 1948 was classified as genotype A, and one isolate from Kenya in 1999 was designated as genotype B. The CVA4 strains segregating with genotype C were collected from Russia during 2006–2014 (n = 14), India in 2010 (n = 8), the United States in 1999 (n = 1) and 2015 (n = 1), Sichuan province, China in 2006 (n = 1) and 2007 (n = 1), Azerbaijan (n = 1) and Turkmenistan (n = 1), respectively. Strikingly, genotype D was the predominant genotype in China. Genotype

[1] http:/beast.bio.ed.ac.uk/Tracer

[2] http:/tree.bio.ed.ac.uk/software/figtree/

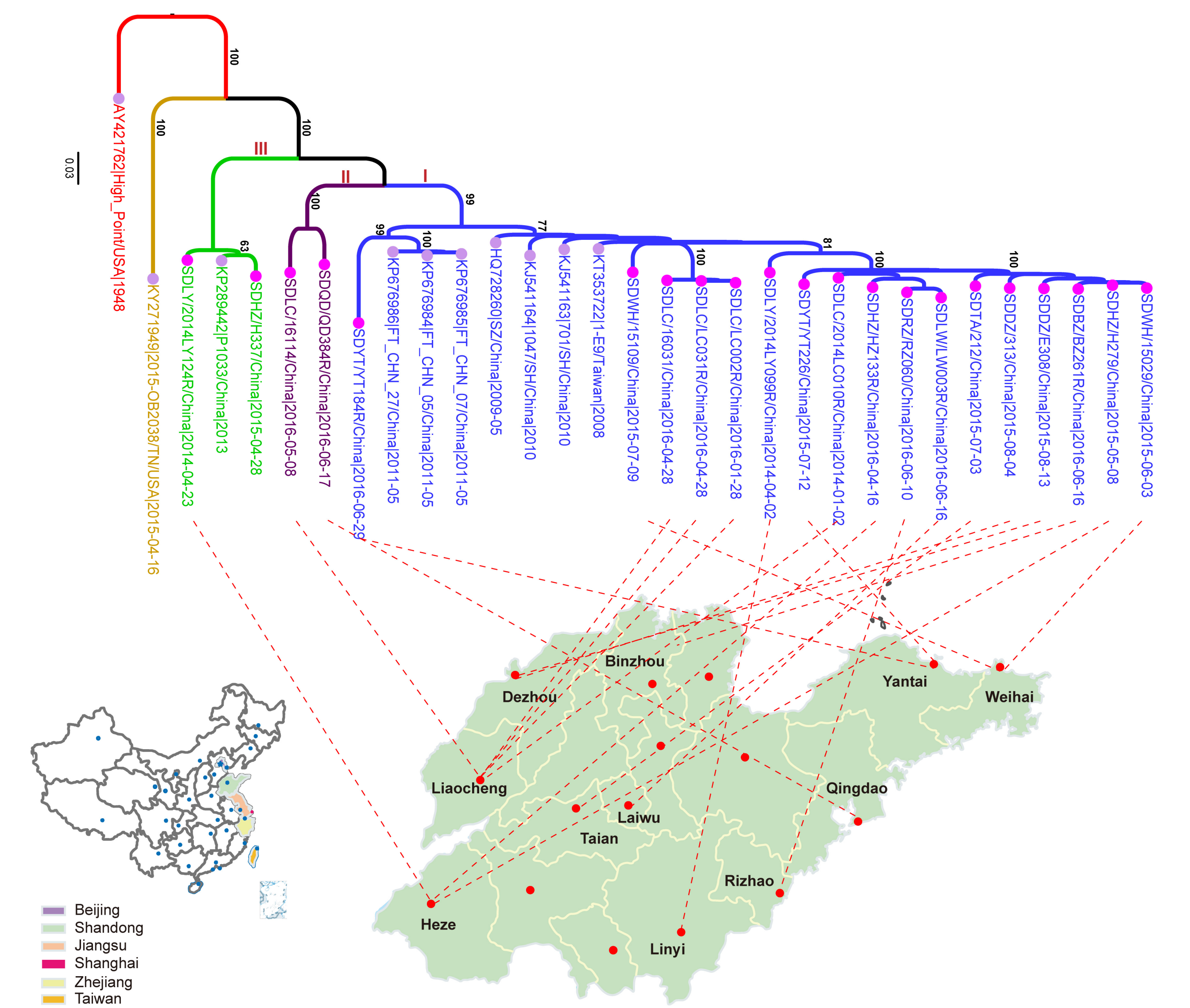

0.03
100
100
100
III
II
I
100
99
99
100
77
63
100
81
100
100
AY421762|High_Point/USA|1948
KY271949|2015-OB2038TN/USA|2015-04-16
SDLY/2014LY124R/China|2014-04-23
KP289442|P1033/China|2013
SDHZ/H337/China|2015-04-28
SDLC/16114/China|2016-05-08
SDQD/QD384R/China|2016-06-17
SDYT/YT184R/China|2016-06-29
KP676986|FT_CHN_27/China|2011-05
KP676984|FT_CHN_05/China|2011-05
KP676985|FT_CHN_07/China|2011-05
HQ728260|SZ/China|2009-05
KJ541164|1047/SH/China|2010
KJ541163|701/SH/China|2010
KT353722|1-E9/Taiwan|2008
SDWH/15109/China|2015-07-09
SDLC/16031/China|2016-04-28
SDLC/LC031R/China|2016-04-28
SDLC/LC002R/China|2016-01-28
SDLY/2014LY099R/China|2014-04-02
SDYT/YT226/China|2015-07-12
SDLC/2014LC010R/China|2014-01-02
SDHZ/HZ133R/China|2016-04-16
SDRZ/RZ060/China|2016-06-10
SDLW/LW003R/China|2016-06-16
SDTA/212/China|2015-07-03
SDDZ/313/China|2015-08-04
SDDZ/E308/China|2015-08-13
SDBZ/BZ261R/China|2016-06-16
SDHZ/H279/China|2015-05-08
SDWH/15029/China|2015-06-03
Binzhou
Dezhou
Yantai
Weihai
Liaocheng
Qingdao
Laiwu
Taian
Rizhao
Heze
Linyi
Beijing
Shandong
Jiangsu
Shanghai
Zhejiang
Taiwan

FIGURE 1 | The maximum likelihood phylogenetic tree of the CVA4 whole genome sequences (*n* = 31). The 21 CVA4 isolates described in this study are colored in rose and the nodes of the reference stains downloaded from GenBank are colored in lavender. The three clades (I, II, and III) are colored in blue, purple and light green, and the prototype strain CVA4 High Point (AY421762) is colored in red.

D could be further divided into two highly supported sub-genotypes: D1 (*n* = 5) and D2 (*n* = 140). In detail, five Chinese strains isolated before 2006 belonged to sub-genotype D1, and the remaining Chinese strains clustered within sub-genotype D2, including the 21 isolates described in the present study. However, our sequenced strains did not cluster together, and were interspersed within sub-genotype D2. Notably, there was one CVA4 strain from Russia, KC879539/40238/Russia/2011, falling within sub-genotype D2, and sharing high homology (97.5%) with KY978552/11142/SD/China/2011.

The mean between group nucleotide and amino acid distances were 21.8 and 2.4% (CVA4 genotype A vs. B), 22.5 and 2.8% (A vs. C), 21.7 and 2.3% (A vs. D), 28.9 and 2.6% (B vs. C), 28.0 and 2.9% (B vs. D), and 20.4 and 3.2% (C vs. D), respectively. In addition, the mean nucleotide (and amino acid) distances between groups D1 and D2 was 15.2% (2.2%). The mean nucleotide (and amino acid) distances within groups C and D were 13.2% (2.8%) and 6.0% (1.3%), respectively, i.e., 15%. The mean nucleotide (and amino acid) distances within groups D1 and D2 were 10.4% (2.0%), and 5.3% (1.2%), respectively.

Pairwise Genetic Diversity Along the CVA4 Genome

Employing a sliding windows analysis across CVA4 genomes (*n* = 31), we found that the pairwise genetic diversity at regions P2 (including the *2A*, *2B*, and *2C* genes) and P3 (including *3A*, *3B*, *3C*, and *3D* genes) was higher than that at region P1 (including

VP4, *VP2*, *VP3*, and *VP1* genes) (**Figure 2A**). In addition, *3A*, *3B* and the 3′ terminus of *3D* showed higher pairwise genetic diversity than other gene regions. The polyprotein gene of the 21 CVA4 genomes from the present study had 82.6–99.2% nucleotide sequence homology and a corresponding 97.1–99.9% amino acid sequence homology, suggesting the occurrence of numerous synonymous mutations. Consistent with this observation, the mean *Sn* at the third position of the codons of the polyprotein gene of the CVA4 strains was 0.224, significantly higher than those at positions 1 and 2 of the codons (**Figure 2B**). Furthermore, the mean *Sn* values at the third position of the codons of *2B*, *2C*, *3A*, *3C,* and *3D* genes were significant higher than those of the *VP1*, *VP2*, *VP3,* and *VP4* genes in the region P1 (**Figure 2C**).

Evolutionary Dynamics of Worldwide CVA4

The Bayesian phylogenetic tree of all CVA4 *VP1* genes ($n = 175$) reconstructed employing the best-fit parameter model revealed the same topology as the maximum likelihood phylogenetic tree (**Figure 3**). The Tempest analysis of molecular clock structure revealed a strong temporal correlation in the *VP1* gene ($R^2 = 0.86$). The mean nucleotide substitution rate for the CVA4 *VP1* genes was estimated to be 5.12×10^{-3} substitutions/site/year (95% highest posterior density (HPD): 4.45×10^{-3}–5.81×10^{-3} substitutions/site/year). Around the year 1941 (95% HPD: 1931–1948), the CVA4 viruses were estimated to diverge into genotype A and the common ancestry of genotypes B, C, and D. Genotype B was suggested to have diverged from the common ancestry around 1951 (95% HPD: 1933–1966), and finally genotypes C and the genotype D diverged around 1970 (95% HPD: 1961–1977). Furthermore, genotype D was further divided into sub-genotypes D1 and D2 which occurred approximately in 1979 (95% HPD: 1974–1985), and the time to the most recent common ancestor (tMRCA) of the sub-genotype D2 could be traced back to 2001 (95% HPD: 1999–2003).

Additionally, we performed a Bayesian skyline plot analysis to reconstruct the demographic history of CVA4 based on the *VP1* gene sequences (**Figure 3**). Our results showed that the effective population size (analog to the number of infected individuals) of CVA4 virus experienced five major stages since the prototype strain High Point was described in 1948 (**Figure 3**). The first stage covered 1948 to 2000 where the population size was relatively constant, with evidence of a slight increase. During this stage, genotypes B, C, and D were already present before 1975, as well as the emergence of genotype D1 in the early 1980s and the subsequent co-circulation of genotypes B, C, and sub-genotype D1. In the second stage, the population size then experienced a sharp and rapid decrease from 2000 to approximately mid-2003 possibly due to a reduction in the prevalence of genotype B and sub-genotype D1 (or potentially a lack of surveillance data), as genotype C and sub-genotype D2 diversified during this period. Soon after mid-2003, the population size of CVA4 experienced a remarkable and rapid growth until 2010, with the widespread circulation of genotype C in Russia and in India, and sub-genotype D2 in China. After 2010, there was another sharp and abrupt decrease in the effective population size which may have been caused by the decreased prevalence of genotype C. Since 2011, CVA4 appears to have entered a rebound stage with the population size becoming stable again with sub-genotype D2 dominating in China.

Molecular Characterizations

Compared with the CVA4 prototype strain High Point, 40 amino acid substitutions were identified in the polyprotein in >75% of our isolates, including 1 in *VP4*, 2 in *VP2*, 2 in *VP3*, 6 in *VP1*, 6 in *2A*, 1 in *2B*, 5 in *2C*, 1 in *3A*, 3 in *3C*, and 11 in *3D* (**Supplementary Table S3**). In EVA71, 14 critical amino acid residues and two gene motifs have been reported previously to be associated with viral infectivity *in vitro* and *in vivo* (**Table 1**) (30–43). Interestingly, the CVA4 strains pocessed the same amino aicd residues at several sites as those of the EVA71 strain BrCr, such as residues 37, 113, and 192 of the *VP1* protein, 84, 82–86, and 154–156 of the *3C* protein, and 73 and 363 of the *3D* protein (**Table 1**). However, different from $VP1_{169L}$ of the prototype EVA71 strain, $VP1_{169F}$ was found in High Point and in each of our 21 CVA4 isolates, indicating that CVA4 may have adapted to bind to the murine scavenger receptor class B member 2 (SCARB2) (Victorio et al., 2016). Notably, there were different amino acid residuals at $VP2_{149}$, $VP1_{97}$, $VP1_{98}$, $VP1_{123}$, $VP1_{167}$, and $VP1_{244}$ between EVA71 BrCr and the CVA4 strains, and $VP1_{145K}$ was observed among our CVA4 isolates.

DISCUSSION

Although EVA71 and CVA16 are known as the major causative agents of HFMD in China, a number of other EV-A species including CVA6, CVA10, CVA2, and CVA4 have been reported to be associated with HFMD[3]. In addition, CVA4 was also one of the common pathogens associated with cases of aseptic meningitis, herpetic angina, and viral myocarditis (Carey and Myers, 1969; Estrada Gonzalez and Mas, 1977; Lee et al., 2014; Wang et al., 2016), which has raised serious public health concern and speculation as to whether there is an altered tropism or pathogenicity associated with this CV type. However, detailed genome analysis has been problematic because there were only ten complete CVA4 genomes available in the GenBank database to date (Hu et al., 2011). Therefore, the molecular epidemiological characterization of CVA4 remains far well less understood due to a paucity of available genomic data.

In this study, we have described 21 novel CVA4 genomes isolated from children with HFMD and revealed that our isolates clustered into three highly supported clades in a maximum likelihood tree estimated using the whole genomes. This suggested that there was a certain genetic diversity in the Chinese CVA4 viruses, which would have been underestimated due to the limited surveillance data available. Phylogenetic analysis of the *VP1* gene sequences showed that the CVA4 strains could be classified into four separable genotypes, A-D,

[3] www.picornaviridae.com

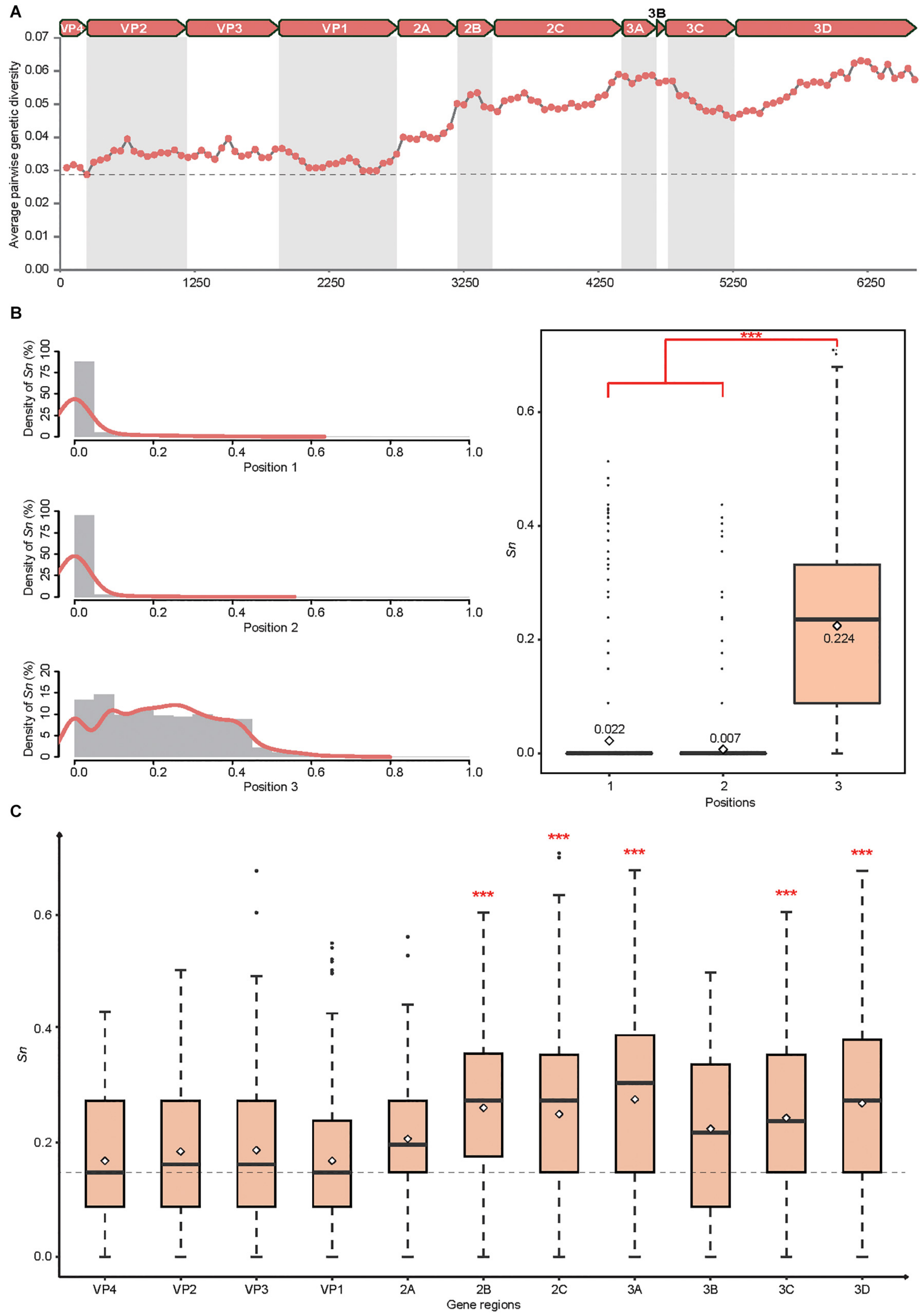


FIGURE 2 | Genetic diversity of the CVA4 polyprotein genes (*n* = 31). **(A)** Average pairwise genetic diversity of all CVA4 polyprotein genes was calculated using a sliding window of 300 nts with a step size of 50 nts. **(B)** Positional entropy values and the *Sn* values at partitions 1, 2, and 3 of each codon were estimated, respectively. ****p*-value <0.001 in the Wilcoxon rank-sum test. **(C)** The *Sn* values of nucleotides at positions three of the codons were estimated in *VP4, VP2, VP3, VP1, 2A, 2B, 2C, 3A, 3B, 3C,* and *3D* genes, respectively. ****p*-value <0.001 in the Wilcoxon rank-sum test between *2B, 2C, 3A, 3C, 3D* genes and *VP4, VP2, VP3, VP1* genes, respectively.

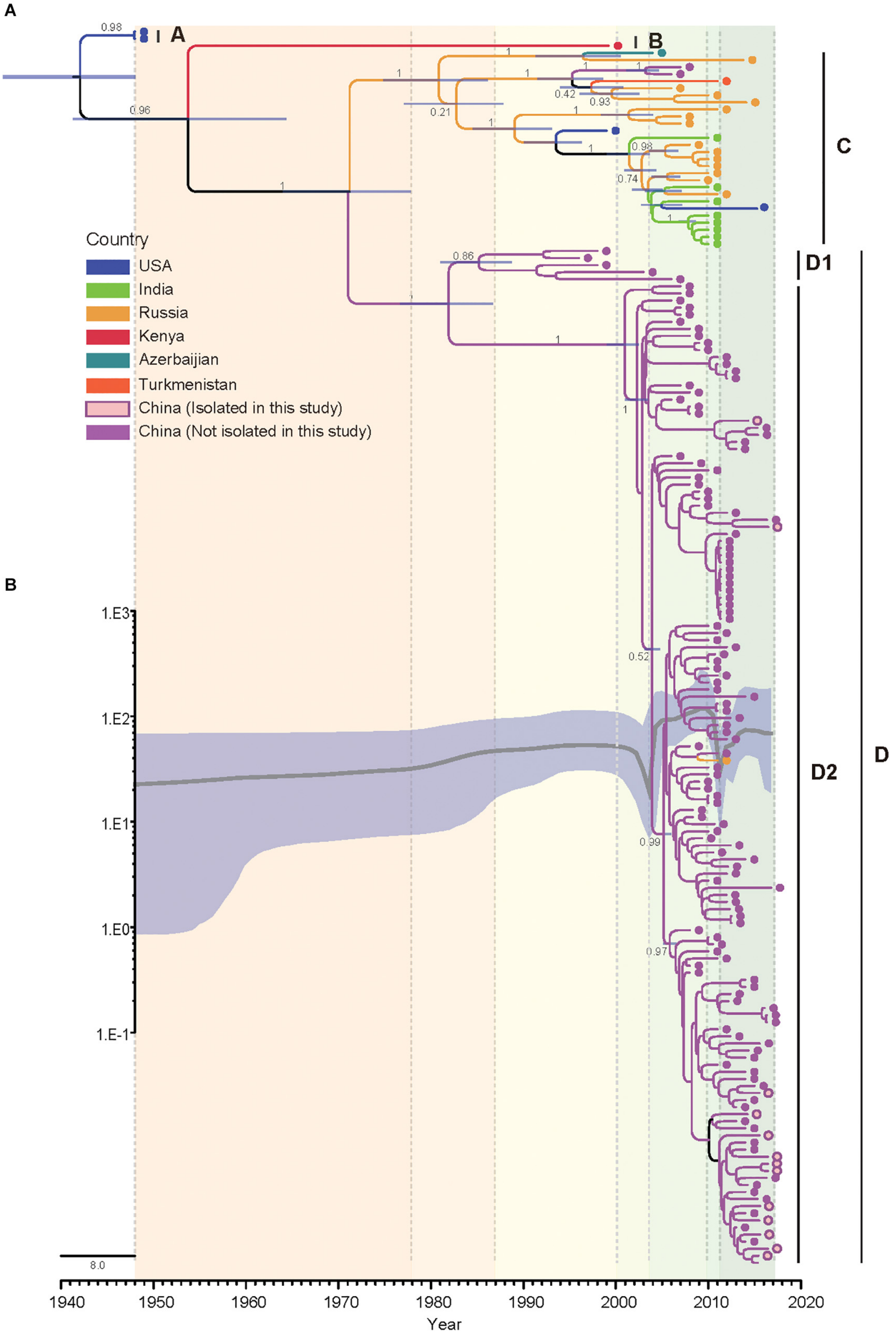


FIGURE 3 | Bayesian phylogenetic analysis and demographic reconstruction of the CVA4 *VP1* gene sequences (915 nts). **(A)** Bayesian phylogeny was conducted by BEAST with a Bayesian skyline tree prior and a relaxed molecular clock model. **(B)** The *x*-axis is the time scale (years) and the *y*-axis is the effective population size in logarithmic Neτ scale. The thick solid line indicates the median estimates and the shaded area indicates the 95% highest posterior density.

TABLE 1 | Comparison of potential critical amino acid residues of the prototype EV71 and CVA4 strains and the 21 isolates described in the present study.

Protein	Amino acid position[a]	EVA71 BrCr (U22521)	CVA4 high point (AY421762)	Shandong isolates ($n = 21$)	Potential biological functions
VP2	L149M	K	S	S	Promoting viral binding and RNA accumulation
					Contributes to viral infectivity *in vitro* and mouse lethality *in vivo* (Huang et al., 2012)
	K149I				Alters the tropism in a receptor-dependent or -independent manner (Miyamura et al., 2011)
					Efficient virus replication in Chinese hamster ovary cells (Chua et al., 2008)
VP1	H37R	H	H	H	Necessary for K244E rescue in primate cell culture (Caine et al., 2016)
	L97R	L	D	D	Confers ability to use HS as an attachment receptor (Tseligka et al., 2018)
	K98E	K	T	T (n = 20), A (n = 1)	Confers binding ability to murine SCARB2 (Victorio et al., 2016)
	I113M	I	I	I	Associated with resistance to pocket-binding compounds NLD, GPP3, ALD (Kelly et al., 2015)
	V123I	V	L	L	
	E145A	R	R	L	Confers binding ability to murine SCARB2 (Victorio et al., 2016)
	E145G/Q				Confers binding ability to PSGL-1 (Nishimura et al., 2013) Important residue for mouse adaptation (Zaini and McMinn, 2012)
	E167G	D	E	E	Stabilizing function based on *VP1 3D* structure (Cordey et al., 2012)
	L169F	L	F	F	Confers binding ability to murine SCARB2 (Victorio et al., 2016)
	V192M	V	V	V	Associate with resistance towards BPR0Z-194 (Shih et al., 2004b)
	K244E	E	N	N (n = 20), D (n = 1)	Important residue for mouse adaptation (Zaini et al., 2012)
					Increased virulence and neuro-tropism in adult interferon-deficient mice (Caine et al., 2016)
3C	R84K	R	R	R	Retains good RNA binding and proteolytic activity of the recombinant *3C* (Shih et al., 2004a)
	KFRDI82-86 deletion or QFQ/KNA	KFRDI	KFRDI	KFRDI	Responsible for RNA binding (Shih et al., 2004a)
	VGK154-156T/SAQ	VGK	VGK	VGK	
3D	Y73H	Y	Y	Y	Resulting in a strong temperature-sensitive phenotype (Arita et al., 2005)
	C363I	C	C	C (n = 20), Y (n = 1)	

[a] *Numbering according to EVA71 (U22521).*

with a mean within genotype genetic distance of 0.00–13.2% and the mean between genotype genetic distance of 20.4–28.9%. Therefore, consistent with previous reports (Rico-Hesse et al., 1987; Hu et al., 2011), different CVA4 genotypes could be defined with >15% of between genotype nucleotide distance in the *VP1* gene. However, in contrast with a previous report (Hu et al., 2011), the prototype strain High Point formed an independent branch as genotype A in our classification. The vast majority of the Chinese strains including the 21 isolates in this study belonged to genotype D, with just two strains identified from Sichuan province in 2006 and 2007 belonging to genotype C. In addition, based on the available evidence sub-genotype D2 appears to have supplanted D1 and become predominant in China over the previous decade.

Further analysis showed that mutations were more likely to occur in the P2 and P3 regions encoding non-structural proteins and at the third position of the codons, consistent with previous reports that there was a general synonymous codon usage pattern for many types of enteroviruses (Liu et al., 2011; Ma et al., 2014; Zhang et al., 2014; Su et al., 2017). Moreover, the *VP1* gene was the most conserved region in our analysis, suggesting that it may be a potential target to design a sensitive and CVA4-specific RT-qPCR diagnostic assay and may also be a candidate epitope for development of vaccines.

Since the mid-1950s, the genetic diversity of CVA4 has gradually increased from 1948 to 1999 and experienced a sharp increase between mid-2003 and 2009, paralleling the increased prevalence of CVA4 in China. However, the population size was estimated to have experienced more substantial declines during 1999 and mid-2003 and from 2009 to 2011 than those in a previous report (Chen et al., 2018). Considering that the precision of demographic reconstruction analysis is sensitive to sampling, pinpointing the abrupt decrease of population size of CVA4 between 1999 and mid-2003 would require further sampling efforts. Therefore, national surveillance of various human enteroviruses causing HFMD is urgently needed to elucidate the complex evolutionary dynamics and co-circulation of the enteroviruses to provide an evidence base for rational diagnostic and prophylactic approaches to mitigate disease burden.

The mean evolutionary rate of the CVA4 *VP1* gene was estimated to be 5.12×10^{-3} substitutions/site/year, slightly lower but of a similar range to 6.4×10^{-3} substitutions/site/year estimated by Chen et al. (2018). We estimated the common ancestor of known CVA4 to have existed approximately 75 years ago, also similar to a previous estimate (71.8 years ago) (Tee et al., 2010). Interestingly, the evolutionary rate of EVA71 *VP1* was estimated to be around 4.6×10^{-3} substitutions/site/year for both EVA71 genogroups B and C (Tee et al., 2010; Chen et al., 2018). The evolutionary rate of CVA6 was estimated at 8.1×10^{-3} substitutions/site/year (Puenpa et al., 2016), and that of CVA16 was 6.7×10^{-3} substitutions/site/year (Zhao et al., 2016). Therefore, CVA4 appears to have been evolving at a similar rate to the other major HFMD pathogens, which deserves further attention.

Coxsackievirus A4 shared phenotype-associated amino acid substitutions when compared to the EVA71 prototype strain, BrCr. Of particular note, the $VP1_{169\text{F}}$ identified in EVA71 may have conferred viral binding ability to the mammalian SCARB2 receptor, also found in the CVA4 isolates (Yamayoshi et al., 2012). In addition, there were several sites with different amino acid motifs between BrCr and the CVA4 viruses. However, it should be noted that although the biological relevance of these sites with specific amino acid substitutions have been confirmed in EVA71, they were identified through sequence comparison between CVA4 and EVA71, Therefore, the effects of these critical amino acids in CVA4 on receptor binding, viral infectivity, replication mammalian adaptation and pathogenesis warrant further studies.

Additionally, we also detected potential genetic recombination events within the CVA4 genomes (n = 31) using the Recombination Detection Program (RDP) v4.16 (Martin et al., 2015) and Simplot v3.5.1 (Hu et al., 2011). However, we failed to detect convincing recombination events in the 21 newly sequenced CVA4 viruses (data not shown). Overall it appears that genetic substitution, rather than recombination, has played a more important role in the evolution and diversification of CVA4, despite recombination events being frequently identified in other human enteroviruses (Hu et al., 2011). However, limited surveillance data may also have led to an underestimation of the number of genetic recombination events on CVA4 evolution.

In summary, we have described 21 novel CVA4 genome sequences, which greatly enriches the available genomic data for CVA4. Phylogenetic analysis revealed that genotypes C, D1 and D2 were co-circulating in the early 2000s and D2 has now supplanted D1 to become the predominant sub-genotype in China. The genetic diversity among Chinese CVA4, occurred at regions encoding the non-structural proteins and at the third wobble position of the codons. In contrast, fewer mutations occurred at the *VP1* gene region, which makes it an optimal candidate region for design of molecular assays and potentially as a conserved epitope for vaccination. Notably, CVA4 may possess higher binding affinity to SCARB2 because of the substitution L169F in the *VP1* protein and how this substitution affects the viral infectivity and clinical outcome should be investigated in the future. In addition to monitoring EVA71 and CVA16, enhanced surveillance of increasingly prevalent and also virulent agents, including CVA4, in China is urgently needed to better understand the dynamics of circulation and evolution of human enteroviruses, which will provide invaluable information for diagnostic and prophylactic approaches to contain disease.

AUTHOR CONTRIBUTIONS

MW, JL, and W-FS designed the study, participated in all tests and drafted the manuscript. M-XY, Y-WZ, S-JD, and X-JW participated in collecting and testing samples. TH, MC, SD, X-CZ, Z-JZ, HZ, and Y-GT conceived the study, contributed to the analysis of the results and preparation of revised manuscript versions. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Technology R&D Program of China (Project No. 2017ZX10104001-006), the Natural Science Foundation of Shandong Province (ZR2018PH034), the Medical and Health Science and Technology Development Projects of Shandong Province (2013WS0157 and 2017WS066), the Tai'an Municipal Science and Technology Development Program (2017NS0107), and the National Natural Science Foundation of China (81601773). W-FS was supported by the Taishan Scholars program of Shandong Province (ts201511056).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01001/full#supplementary-material

REFERENCES

Arita, M., Shimizu, H., Nagata, N., Ami, Y., Suzaki, Y., Sata, T., et al. (2005). Temperature-sensitive mutants of enterovirus 71 show attenuation in cynomolgus monkeys. *J. Gen. Virol.* 86(Pt 5), 1391–1401. doi: 10.1099/vir.0.80784-0

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., and Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* 29, 2157–2167. doi: 10.1093/molbev/mss084

Cai, C., Yao, X., Zhuo, F., He, Y., and Yang, G. (2015). [Gene characterization of the VP1 region of Coxsackievirus A4 from herpangina cases in Shenzhen of China]. *Zhonghua Yi Xue Za Zhi* 95, 1226–1229.

Caine, E. A., Moncla, L. H., Ronderos, M. D., Friedrich, T. C., and Osorio, J. E. (2016). A single mutation in the VP1 of enterovirus 71 Is responsible for increased virulence and neurotropism in adult interferon-deficient mice. *J. Virol.* 90, 8592–8604. doi: 10.1128/JVI.01370-16

Carey, D. E., and Myers, R. M. (1969). Isolation of type A4 coxsackie virus from the blood serum of a child with rapidly fatal illness marked by severe central nervous system manifestations. *Indian J. Med. Res.* 57, 765–769.

Chen, P., Wang, H., Tao, Z., Xu, A., Lin, X., Zhou, N., et al. (2018). Multiple transmission chains of coxsackievirus A4 co-circulating in China and neighboring countries in recent years: phylogenetic and spatiotemporal analyses based on virological surveillance. *Mol. Phylogenet. Evol.* 118, 23–31. doi: 10.1016/j.ympev.2017.09.014

Chu, P. Y., Lu, P. L., Tsai, Y. L., Hsi, E., Yao, C. Y., Chen, Y. H., et al. (2011). Spatiotemporal phylogenetic analysis and molecular characterization of coxsackievirus A4. *Infect. Genet. Evol.* 11, 1426–1435. doi: 10.1016/j.meegid.2011.05.010

Chua, B. H., Phuektes, P., Sanders, S. A., Nicholls, P. K., and McMinn, P. C. (2008). The molecular basis of mouse adaptation by human enterovirus 71. *J. Gen. Virol.* 89(Pt 7), 1622–1632. doi: 10.1099/vir.0.83676-0

Cordey, S., Petty, T. J., Schibler, M., Martinez, Y., Gerlach, D., van Belle, S., et al. (2012). Identification of site-specific adaptations conferring increased neural cell tropism during human enterovirus 71 infection. *PLoS Pathog.* 8:e1002826. doi: 10.1371/journal.ppat.1002826

Dalldorf, G. (1953). The coxsackie virus group. *Ann. N. Y. Acad. Sci.* 56, 583–586. doi: 10.1111/j.1749-6632.1953.tb30251.x

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9:772. doi: 10.1038/nmeth.2109

Drummond, A. J., and Rambaut, A. (2007). BEAST: bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214. doi: 10.1186/1471-2148-7-214

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

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Estrada Gonzalez, R., and Mas, P. (1977). Virological studies in acute polyradiculoneuritis, LGBS type. Various findings in relation to Coxsackie A4 virus. *Neurol. Neurocir. Psiquiatr.* 18(2–3 Suppl.), 527–531.

Faria, N. R., Quick, J., Claro, I. M., Theze, J., de Jesus, J. G., Giovanetti, M., et al. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* 546, 406–410. doi: 10.1038/nature22401

Hu, Y. F., Yang, F., Du, J., Dong, J., Zhang, T., Wu, Z. Q., et al. (2011). Complete genome analysis of coxsackievirus A2, A4, A5, and A10 strains isolated from hand, foot, and mouth disease patients in China revealing frequent recombination of human enterovirus A. *J. Clin. Microbiol.* 49, 2426–2434. doi: 10.1128/JCM.00007-11

Huang, S. W., Wang, Y. F., Yu, C. K., Su, I. J., and Wang, J. R. (2012). Mutations in VP2 and VP1 capsid proteins increase infectivity and mouse lethality of enterovirus 71 by virus binding and RNA accumulation enhancement. *Virology* 422, 132–143. doi: 10.1016/j.virol.2011.10.015

Kelly, J. T., De Colibus, L., Elliott, L., Fry, E. E., Stuart, D. I., Rowlands, D. J., et al. (2015). Potent antiviral agents fail to elicit genetically-stable resistance mutations in either enterovirus 71 or Coxsackievirus A16. *Antivir. Res.* 124, 77–82. doi: 10.1016/j.antiviral.2015.10.006

Knowles, N. J., Hovi, T., Hyypiä, T., King, A. M. Q., Lindberg, A. M., Pallansch, M. A., et al. (2012). "Picornaviridae," 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. Lefkowitz, and E. B. Carstens (San Diego, CA: Elsevier).

Lee, C. J., Huang, Y. C., Yang, S., Tsao, K. C., Chen, C. J., Hsieh, Y. C., et al. (2014). Clinical features of coxsackievirus A4, B3 and B4 infections in children. *PLoS One* 9:e87391. doi: 10.1371/journal.pone.0087391

Li, J. S., Dong, X. G., Qin, M., Xie, Z. P., Gao, H. C., Yang, J. Y., et al. (2015). Outbreak of febrile illness caused by coxsackievirus A4 in a nursery school in Beijing, China. *Virol. J.* 12:92. doi: 10.1186/s12985-015-0325-1

Liu, J. F., Zhang, Y., and Li, H. (2009). [Genetic characterization of VP4-VP2 of two coxsackievirus A4 isolated from patients with hand, foot and mouth disease]. *Zhongguo Yi Miao He Mian Yi* 15, 345–349.

Liu, Y. S., Zhou, J. H., Chen, H. T., Ma, L. N., Pejsak, Z., Ding, Y. Z., et al. (2011). The characteristics of the synonymous codon usage in enterovirus 71 virus and the effects of host on the virus in codon usage pattern. *Infect. Genet. Evol.* 11, 1168–1173. doi: 10.1016/j.meegid.2011.02.018

Ma, M. R., Hui, L., Wang, M. L., Tang, Y., Chang, Y. W., Jia, Q. H., et al. (2014). Overall codon usage pattern of enterovirus 71. *Genet. Mol. Res.* 13, 336–343. doi: 10.4238/2014.January.21.1

Martin, D. P., Murrell, B., Golden, M., Khoosal, A., and Muhire, B. (2015). RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol.* 1:vev003. doi: 10.1093/ve/vev003

Melnick, J. L. (1950). Studies on the Coxsackie viruses; properties, immunological aspects and distribution in nature. *Bull. N. Y. Acad. Med.* 26, 342–356.

Mine, I., Taguchi, M., Sakurai, Y., and Takeuchi, M. (2017). Bilateral idiopathic retinal vasculitis following coxsackievirus A4 infection: a case report. *BMC Ophthalmol.* 17:128. doi: 10.1186/s12886-017-0523-2

Miyamura, K., Nishimura, Y., Abo, M., Wakita, T., and Shimizu, H. (2011). Adaptive mutations in the genomes of enterovirus 71 strains following infection of mouse cells expressing human P-selectin glycoprotein ligand-1. *J. Gen. Virol.* 92(Pt 2), 287–291. doi: 10.1099/vir.0.022418-0

Nishimura, Y., Lee, H., Hafenstein, S., Kataoka, C., Wakita, T., Bergelson, J. M., et al. (2013). Enterovirus 71 binding to PSGL-1 on leukocytes: VP1-145 acts as a molecular switch to control receptor interaction. *PLoS Pathog.* 9:e1003511. doi: 10.1371/journal.ppat.1003511

Oberste, M. S., Penaranda, S., Maher, K., and Pallansch, M. A. (2004). Complete genome sequences of all members of the species Human enterovirus A. *J. Gen. Virol.* 85(Pt 6), 1597–1607. doi: 10.1099/vir.0.79789-0

Palacios, G., Casas, I., Tenorio, A., and Freire, C. (2002). Molecular identification of enterovirus by analyzing a partial VP1 genomic region with different methods. *J. Clin. Microbiol.* 40, 182–192. doi: 10.1128/jcm.40.1.182-192.2002

Puenpa, J., Vongpunsawad, S., Osterback, R., Waris, M., Eriksson, E., Albert, J., et al. (2016). Molecular epidemiology and the evolution of human coxsackievirus A6. *J. Gen. Virol.* 97, 3225–3231. doi: 10.1099/jgv.0.000619

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* 2:vew007. doi: 10.1093/ve/vew007

Ramirez, C., Gregori, J., Buti, M., Tabernero, D., Camos, S., Casillas, R., et al. (2013). A comparative study of ultra-deep pyrosequencing and cloning to quantitatively analyze the viral quasispecies using hepatitis B virus infection as a model. *Antivir. Res.* 98, 273–283. doi: 10.1016/j.antiviral.2013.03.007

Rico-Hesse, R., Pallansch, M. A., Nottay, B. K., and Kew, O. M. (1987). Geographic distribution of wild poliovirus type 1 genotypes. *Virology* 160, 311–322. doi: 10.1016/0042-6822(87)90001-8

Shih, S. R., Chiang, C., Chen, T. C., Wu, C. N., Hsu, J. T., Lee, J. C., et al. (2004a). Mutations at KFRDI and VGK domains of enterovirus 71 3C protease affect its RNA binding and proteolytic activities. *J. Biomed. Sci.* 11, 239–248. doi: 10.1159/000076036

Shih, S. R., Tsai, M. C., Tseng, S. N., Won, K. F., Shia, K. S., Li, W. T., et al. (2004b). Mutation in enterovirus 71 capsid protein VP1 confers resistance to the inhibitory effects of pyridyl imidazolidinone. *Antimicrob. Agents Chemother.* 48, 3523–3529. doi: 10.1128/AAC.48.9.3523-3529.2004

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456–463. doi: 10.1093/bioinformatics/bti191

Su, W., Li, X., Chen, M., Dai, W., Sun, S., Wang, S., et al. (2017). Synonymous codon usage analysis of hand, foot and mouth disease viruses: a comparative study on coxsackievirus A6, A10, A16, and enterovirus 71 from 2008 to 2015. *Infect. Genet. Evol.* 53, 212–217. doi: 10.1016/j.meegid.2017.06.004

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28, 2731–2739. doi: 10.1093/molbev/msr121

Tee, K. K., Lam, T. T., Chan, Y. F., Bible, J. M., Kamarulzaman, A., Tong, C. Y., et al. (2010). Evolutionary genetics of human enterovirus 71: origin, population dynamics, natural selection, and seasonal periodicity of the VP1 gene. *J. Virol.* 84, 3339–3350. doi: 10.1128/JVI.01019-09

Tseligka, E. D., Sobo, K., Stoppini, L., Cagno, V., Abdul, F., Piuz, I., et al. (2018). A VP1 mutation acquired during an enterovirus 71 disseminated infection confers heparan sulfate binding ability and modulates ex vivo tropism. *PLoS Pathog.* 14:e1007190. doi: 10.1371/journal.ppat.1007190

Ueda, Y., Kenzaka, T., Noda, A., Yamamoto, Y., and Matsumura, M. (2015). Adult-onset Kawasaki disease (mucocutaneous lymph node syndrome) and concurrent Coxsackievirus A4 infection: a case report. *Int. Med. Case Rep. J.* 8, 225–230. doi: 10.2147/IMCRJ.S90685

Victorio, C. B., Xu, Y., Ng, Q., Meng, T., Chow, V. T., and Chua, K. B. (2016). Cooperative effect of the VP1 amino acids 98E, 145A and 169F in the productive infection of mouse cell lines by enterovirus 71 (BS strain). *Emerg. Microbes Infect.* 5:e60. doi: 10.1038/emi.2016.56

Wang, D., Xu, Y., Zhang, Y., Zhu, S., Si, Y., Yan, D., et al. (2016). [Genetic characteristics of coxsackievirus group A type 4 isolated from patients with acute flaccid paralysis in Shaanxi, China]. *Bing Du Xue Bao* 32, 145–149.

Wang, J., Hu, T., Sun, D., Ding, S., Carr, M. J., Xing, W., et al. (2017). Epidemiological characteristics of hand, foot, and mouth disease in Shandong, China, 2009-2016. *Sci. Rep.* 7:8900. doi: 10.1038/s41598-017-y, we performed a Bayesian sk09196-z

Yamayoshi, S., Iizuka, S., Yamashita, T., Minagawa, H., Mizuta, K., Okamoto, M., et al. (2012). Human SCARB2-dependent infection by coxsackievirus A7, A14, and A16 and enterovirus 71. *J. Virol.* 86, 5686–5696. doi: 10.1128/JVI.00020-12

Zaini, Z., and McMinn, P. (2012). A single mutation in capsid protein VP1 (Q145E) of a genogroup C4 strain of human enterovirus 71 generates a mouse-virulent phenotype. *J. Gen. Virol.* 93(Pt 9), 1935–1940. doi: 10.1099/vir.0.043893-0

Zaini, Z., Phuektes, P., and McMinn, P. (2012). Mouse adaptation of a sub-genogroup B5 strain of human enterovirus 71 is associated with a novel lysine to glutamic acid substitution at position 244 in protein VP1. *Virus Res.* 167, 86–96. doi: 10.1016/j.virusres.2012.04.009

Zhang, H., Cao, H. W., Li, F. Q., Pan, Z. Y., Wu, Z. J., Wang, Y. H., et al. (2014). Analysis of synonymous codon usage in enterovirus 71. *Virusdisease* 25, 243–248. doi: 10.1007/s13337-014-0215-y

Zhang, Z., Dong, Z., Li, J., Carr, M. J., Zhuang, D., Wang, J., et al. (2017a). Protective efficacies of formaldehyde-inactivated whole-virus vaccine and antivirals in a murine model of coxsackievirus a10 infection. *J. Virol.* 91:e00333-17. doi: 10.1128/JVI.00333-17

Zhang, Z., Dong, Z., Wei, Q., Carr, M. J., Li, J., Ding, S., et al. (2017b). A neonatal murine model of coxsackievirus A6 infection for evaluation of antiviral and vaccine efficacy. *J. Virol.* 91:e02450-16. doi: 10.1128/JVI.02450-16

Zhao, G., Zhang, X., Wang, C., Wang, G., and Li, F. (2016). Characterization of VP1 sequence of Coxsackievirus A16 isolates by Bayesian evolutionary method. *Virol. J.* 13:130. doi: 10.1186/s12985-016-0578-3

Zhen, R., Zhang, Y., Xie, H., Chen, C., Geng, J., He, P., et al. (2014). [Sequence analysis of VP1 region of coxsackievirus A4 and coxsackievirus A10 in Guangzhou city, 2010-2012]. *Zhonghua Yu Fang Yi Xue Za Zhi* 48, 445–450.

Conflict of Interest Statement: 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.

TECHNOLOGY REPORT
published: 08 May 2019
doi: 10.3389/fmicb.2019.01056

A Mismatch-Tolerant Reverse Transcription Loop-Mediated Isothermal Amplification Method and Its Application on Simultaneous Detection of All Four Serotype of Dengue Viruses

Yi Zhou[1,2†], Zhenzhou Wan[3†], Shuting Yang[4†], Yingxue Li[2], Min Li[5], Binghui Wang[4], Yihong Hu[2], Xueshan Xia[4], Xia Jin[5], Na Yu[1] and Chiyu Zhang[2]**

[1] School of Life Sciences, East China Normal University, Shanghai, China, [2] Pathogen Discovery and Big Data Center, Chinese Academy of Sciences (CAS) Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China, [3] Medical Laboratory of Taizhou Fourth People's Hospital, Taizhou, China, [4] Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming, China, [5] Viral Disease and Vaccine Translational Research Unit, CAS Key Laboratory of Molecular Virology and Immunology, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana-Champaign, United States

Reviewed by:
Kim-Kee Tan,
University of Malaya, Malaysia
Sing-Sin Sam,
University of Malaya, Malaysia
Boon Teong Teoh,
University of Malaya, Malaysia

****Correspondence:***
Na Yu
nyu@bio.ecnu.edu.cn
Chiyu Zhang
zhangcy1999@ips.ac.cn

† These authors have contributed equally to this work

Received: *03 January 2019*
Accepted: *26 April 2019*
Published: *08 May 2019*

Citation:
Zhou Y, Wan Z, Yang S, Li Y, Li M, Wang B, Hu Y, Xia X, Jin X, Yu N and Zhang C (2019) A Mismatch-Tolerant Reverse Transcription Loop-Mediated Isothermal Amplification Method and Its Application on Simultaneous Detection of All Four Serotype of Dengue Viruses. Front. Microbiol. 10:1056. doi: 10.3389/fmicb.2019.01056

Loop-mediated isothermal amplification (LAMP) has been widely used in the detection of pathogens causing infectious diseases. However, mismatches between primers (especially in the 3′-end) and templates significantly reduced the amplification efficiency of LAMP, and limited its application to genetically diverse viruses. Here, we reported a novel mismatch-tolerant LAMP assay and its application in the detection of dengue viruses (DENV). The novel method features the addition of as little as 0.15 U of high-fidelity DNA polymerase to the standard 25 μl LAMP reaction mixture. This amount was sufficient to remove the mismatched bases at the 3′-end of primers, thereby resulting in excellent tolerance for various mismatches occurring at the 3′-end of the LAMP primers during amplification. This novel LAMP assay has a markedly improved amplification efficiency especially for the mutants forming mismatches with internal primers (FIP/BIP) and loop primers (FLP/BLP). The reaction time of the novel method was about 5.6–22.6 min faster than the conventional LAMP method regardless of the presence or absence of mismatches between primers and templates. Using the novel method, we improved a previously established pan-serotype assay for DENV, and demonstrated greater sensitivity for detection of four DENV serotypes than the previous one. The limit of detection (LOD) of the novel assay was 74, 252, 78, and 35 virus RNA copies per reaction for DENV-1, DENV-2, DENV-3, and DENV-4, respectively. Among 153 clinical samples from patients with suspected DENV infection, the novel assay detected 94.8% samples being DENV positive, higher than that detected by the commercial NS1 antigen assay (92.2%), laboratory-based RT-PCR method (78.4%), and the conventional RT-LAMP assay (86.9%). Furthermore, the novel RT-LAMP assay has been developed into a visual determination method by adding colorimetric dyes. Because of its simplicity, all LAMP-based diagnostic assays may be easily updated to the newly improved version. The novel mismatch-tolerant LAMP method represents a simple, sensitive and promising approach for molecular diagnosis of highly variable viruses, and it is especially suited for application in resource-limited settings.

Keywords: mismatch-tolerant LAMP, dengue virus, virus variant, high-fidelity DNA polymerase, mismatch, serotype

INTRODUCTION

Because of their high sensitivity and specificity, nucleic acid amplification (NAA) tests have been explored as molecular diagnosis tools for infectious diseases, especially for acute infections (Saijo et al., 2008; de Paz et al., 2014). There are two main types of NAA techniques: thermal cycling and isothermal amplification methods (de Paz et al., 2014). Thermal cycling amplification is the basis of various PCR-based techniques, such as the quantitative PCR (qPCR) method that has been widely applied to biomedical research, as well as agricultural, ecological, food and environmental sciences (Bustin, 2000, 2002; Saijo et al., 2008; Smith and Osborn, 2009; Kralik and Ricchi, 2017). In particular, a large number of commercial kits were developed using various qPCR techniques. However, PCR-based methods, especially qPCR, require relatively sophisticated equipment and highly trained personnel, to be performed in special diagnostic laboratories or facilities. These requirements limit their applications in resource-limited settings (Yang and Rothman, 2004).

Distinct from PCR-based methods, isothermal amplification is performed under a constant temperature, and rarely relies on sophisticated equipment (Yan et al., 2014; Zhao et al., 2015). Therefore, isothermal amplification techniques represent a promising direction for the development of point-of-care testing (POCT) method for use in the fields or resource-limited settings (de Paz et al., 2014). Some isothermal amplification methods have already been developed, such as nucleic acid sequence-based amplification (NASBA), loop mediated isothermal amplification (LAMP), rolling circle amplification (RCA), and recombinase polymerase amplification (RPA) (Yan et al., 2014; Zhao et al., 2015). Among them, LAMP is the most commonly used technique with more than 3,000 PubMed searchable publications by January, 2019 (Notomi et al., 2000). LAMP generally uses six primers to initiate the self-primed DNA synthesis and amplification cycling (Notomi et al., 2000; Nagamine et al., 2002). Although the most conserved genomic region is generally used for primer design, the LAMP primers, especially the two inner LAMP primers FIP and BIP that are typically over 40 nt long, form mismatches easily with templates of highly variable viruses that exist as a quasispecies (Sanjuan et al., 2010; Domingo and Perales, 2018). The mismatches between primers and templates can significantly reduce the amplification efficiency of LAMP (Notomi et al., 2000), decrease the sensitivity of detection, and even generate false negative results. These drawbacks are a major barrier to translate the LAMP technique into a commercial viable application (Wong et al., 2018), and the reason why few LAMP-based commercial diagnostic kits had been approved by the Food and Drug Administration (FDA) of US or World Health Organization (WHO) for the diagnosis of infectious diseases.

We recently developed a mismatch-tolerant RT-qPCR method with proof-reading capacity by introducing a small amount of high-fidelity DNA polymerase to a standard amount of Taq DNA polymerase (Li et al., 2019). The high-fidelity DNA polymerase removed the mismatched bases at 3′end of the primer and thus significantly improved the amplification efficiency for template containing mutations. Here, we utilized a similar principle to develop a mismatch-tolerant LAMP method for high-sensitive and broad-spectrum detection of genetically diverse viruses, and used dengue virus (DENV) in the proof-of-principle experiments.

DENV was chosen in part because it is a global public health problem with an estimated 390 million infections occurring each year (Bhatt et al., 2013). DENV is a positive-sense single-stranded RNA virus that belongs to the *Flaviviridae* family (Henchal and Putnak, 1990). Dengue epidemics occur mainly in urban and semi-urban areas in the tropical and subtropical regions, including a large number of resource-poor countries (Bhatt et al., 2013). The diseases caused by DENV infection vary from mild dengue fever to life-threatening dengue hemorrhagic fever or dengue shock syndrome (Henchal and Putnak, 1990; Gubler, 1998; Ligon, 2005). Symptomatic DENV infections are readily recognizable by clinical manifestation, serological and molecular diagnostic assays. However, the majority of individuals with DENV infection are subclinical, with no manifested symptoms, yet capable of spreading the virus (Bhatt et al., 2013). Therefore, early and accurate diagnosis of DENV infection, using a relatively simple method that can be deployed to different terrains, is critical not only to clinical management, but also for the prevention and control of dengue epidemic (Teles et al., 2005).

Part of the difficulty in dengue diagnosis is that DENV comprises four antigenically distinct serotypes of DENV-1, DENV-2, DENV-3, and DENV-4, and different serotypes only exhibit 65–70% sequence homology at the nucleotide level (Henchal and Putnak, 1990; Ligon, 2005). Although some detection assays, including antibody detection (IgM and IgG), antigen detection (non-structural protein 1, NS1), and various NAA assays for individual or multiple serotypes, had been developed (Shu and Huang, 2004; Teles et al., 2005), there still lacks a high-efficiency pan-serotype DENV detection assay for point-of-care use in resource-limited settings. Using the novel mismatch-tolerant

LAMP technique, we improved the sensitivity of detection of a previous developed pan-serotype RT-LAMP assay for DENV, and demonstrated that such a test may have practical applications.

MATERIALS AND METHODS

Viruses

Four DENV strains DENV-1 (16007), DENV-2 (16681), DENV-3 (16562), and DENV-4 (1036) were used to establish the novel DENV RT-LAMP assay. Two other flaviviruses, Zika virus (SZWIV-01) and attenuated yellow fever virus 17D (derived from Asibi strain), and eight respiratory viruses were used to evaluate the cross-reactivity of the assay. The respiratory viruses include adenovirus (VR-930), enterovirus (VR-1076), influenza A and B viruses (VR-333 and VR-789), parainfluenza viruses 3 (VR-93), HCoV-229E (VR-740), HCoV-OC43 (VR-1558), and human rhinovirus (VR-489). Four DENV strains were gifts from Dr. Claire Huang (U.S Centers for Disease Control and Prevention at Fort Collins, Colorado). The Asian ZIKV strain SZ-WIV01, isolated from human serum of an imported case of ZIKV infection in China in 2016, was obtained from Wuhan Institute of Virology, Chinese Academy of Sciences. The yellow fever virus 17D was obtained from Dr. Cheng-feng Qin (Academy of Military Medical Sciences, Beijing). The respiratory viruses were previously purchased from the American Type Culture Collection (ATCC) and stored at Institut Pasteur of Shanghai, Chinese Academy of Sciences.

RNA Extraction

Viral RNA was extracted from 140 μl virus culture supernatants or patients' plasma using the QIAamp Viral Min Kit (Qiagen, Germany). RNA was eluted in 50 μl nuclease-free water and stored at −80°C until use.

Construction of Mutant Viral RNA Template

To perform the proof-of-concept experiment of the mismatch-tolerant RT-LAMP, a series of mutants that form mismatches with the F3, FIP and BLP primers were constructed using a fast mutagenesis system (TransGen, Beijing, China) based on the 3′-UTR region of the DENV-4 genome. In brief, the 3′-UTR fragment of DENV-4 was obtained using StarScript II Probe One-Step qRT-PCR Kit (GenStar, Beijing, China) with primers F3/134 and B3/4, and then sub-cloned into pMD18-T plasmid vector (TaKaRa, Dalian, China). A T7 promoter was fused to the 5′-end of the primer F3/134. Various mutants were constructed using site-directed mutagenesis with mutagenic primers (**Supplementary Table S1**) and confirmed by Sanger sequencing. Mutant RNAs were obtained through *in vitro* transcription with T7 RNA polymerase, and quantified using spectrophotometric absorbance at 260 nm (NanoDrop, Technologies Inc.). Copy number of each mutant RNA was calculated using the following formula: RNA copies/μl = [RNA concentration (g/μl)/(nt transcript length × 340)] × 6.022×10^{23}.

Reaction System of the Novel RT-LAMP Assay

Bst 2.0 DNA polymerase, WarmStart® RT and Q5 High-Fidelity DNA polymerase (all from New England Biolabs, Beverly, MA, United States) were used to establish the novel mismatch-tolerant RT-LAMP system. The main difference between the novel and the conventional RT-LAMP assays is the inclusion of an additional amount of high-fidelity DNA polymerase in the reaction system. The amount of the Q5 high-fidelity DNA polymerase was optimized at 0.15 U per 25 μl reaction mix. A 25 μl reaction mix of the novel RT-LAMP assay includes 1× isothermal amplification buffer, 6 mM MgSO4, 1.4 mM dNTPs, 8 units of Bst 2.0 DNA polymerase, 7.5 units of WarmStart® RT, 0.15 unit of Q5 High-Fidelity DNA Polymerase, 20 pmol each of primers FIP/123, FIP/4, BIP/123, and BIP/4, 2.5 pmol each of primers F3/134, F3/2, B3/123, and B3/4, and 20 pmol of loop primer BLP/1234. Three microliter of RNA was used for the RT-LAMP assays. The primer information was previously described and provided as **Supplementary Table S1** (Teoh et al., 2013). The RT-LAMP reaction was performed at 63°C for 60 min with 0.4 μM SYTO 9 (Life technologies, Carlsbad, CA, United States) as fluorescent dye for real-time monitoring by the Light Cycler 96 real-time PCR System (Roche Diagnostics, Mannheim, Germany).

Sensitivity and Limit of Detection (LOD)

Ten-fold serial dilution of the RNA from four DENV stocks were used as the standards to determine the sensitivity of the novel DENV RT-LAMP assay. The initial titers of the DENV-1, DENV-2, DENV-3, and DENV-4 stocks were 4×10^6, 1×10^7, 2×10^4, and 1.6×10^7 plaque-forming unit (PFU) per ml. To determine the RNA copy numbers of the virus stocks, RT-qPCR assay was performed with previously described primers (**Supplementary Table S1**; Go et al., 2016), and a standard curve was obtained by 10-fold serial dilutions of *in vitro* transcribed DENV RNA standard from 10^7 to 10^2 copies/μl. The RNA copy numbers of DENV-1, DENV-2, DENV-3, and DENV-4 stocks were quantified to be 1.6×10^8, 1.1×10^8, 9.3×10^6, and 1.2×10^7 copies/ml, respectively.

LOD of the novel DENV RT-LAMP assay was determined using 10-fold serial dilutions of RNA from each virus stock. Each dilution was tested in a set of 10 replicates. To more accurately estimate the LOD, we performed additional experiments using a fivefold dilution of each DENV serotype with a starting RNA copy number of 3×10^4. Probit regression analysis was performed to determine the LOD using the SPSS 17.0 software. The LOD was defined as a 95% probability of obtaining a positive result (Anderson, 1989).

Evaluation of the Novel DENV Detection Assay Using Clinical Samples

To evaluate the performance of the novel RT-LAMP assay for pan-serotype detection of DENV, 153 plasma samples that were previously collected from dengue-suspected patients (Wang et al., 2016, 2018). DENV infection was determined by both NS1 antigen detection and specific RT-PCR assays. The NS1 antigen assay was performed using One-Step Dengue NS1 RapiDip^{TM} InstaTest kit (Cortez Diagnostics, United States); the RT-PCR assay was performed using PrimeScript^{TM} One Step RT-PCR Kit Ver.2 (Takara, Japan). Of the 153 samples, 120 positive samples by RT-PCR assay were further subjected to Sanger sequencing and phylogenetic analysis, and 58 DENV-1, 5 DENV-2, 46 DENV-3, and 11 DENV-4 were identified. During September, 2018, the RNA stocks were tested again using both the novel and the conventional RT-LAMP assays. A comparison in performance of the four methods was performed, and the concordance rate was calculated by Clinical Calculator with the following formula: (number of consistent results by both methods/total number) × 100%[1].

The samples having a negative result by RT-PCR assay but positive by the novel RT-LAMP assay, or having a time threshold (Tt) difference of more than 15 min between the novel and the conventional RT-LAMP assays were subjected to further RT-nested PCR amplifications. The amplicons were processed for Sanger sequencing and further sequence analyses. The RT-nested PCR assay was performed using StarScript II Probe One-Step qRT-PCR Kit (GenStar, Beijing, China) with F3 and B3 as outer primers, and F2 and B2 as inner primers (**Supplementary Table S1**).

Visual Detection

To develop a visual detection method, 120 μM hydroxynaphthol blue (HNB) (Sigma-Aldrich, United States) was added to the novel and the conventional RT-LAMP reaction mix. Another visual RT-LAMP assay was performed with the WarmStart Colorimetric LAMP 2X Master Mix (New England Biolabs, Beverly, MA, United States), which uses cresol red as a visual indicator. The reactions were performed at 63°C, and the color change was observed at the 30-, 40-, 50-, and 60-min time points.

Ethics Statement

This study was approved by the Ethics Committees of Shanghai Public Health Clinical Center (No. 2014-008) and Kunming University of Science and Technology (No. 2018JC027). Written informed consents were obtained from all dengue-suspected patients.

RESULTS

The Establishment of a Mismatch-Tolerant RT-LAMP Assay

The principle of LAMP shows that the 3′-ends of the F2/B2, F3/B3, and LF/LB, as well as the 3′-ends of F1/B1 (corresponding to the 5′-ends of the F1c/B1c) provide 3′-hydroxyl groups for DNA extension (Notomi et al., 2000). Because of lacking a 3′–5′ exonuclease activity, Bst2.0 DNA polymerase has a reduced capacity to extend primers when mismatches occur at the 3′-ends of these primers, and thus resulted in a low amplification efficiency of LAMP (**Figure 1A**). Removing the mismatched bases at the 3′-end of the primers by an additional enzyme can overcome the inhibition effect of mismatches on LAMP reaction.

High-fidelity DNA polymerase has a proofreading function and can be used to remove mismatched nucleotides. To improve the amplification efficiency of LAMP on highly variable templates, we developed a mismatch-tolerant LAMP method by combining Bst 2.0 DNA polymerase and a small amount of high-fidelity DNA polymerase in the reaction (**Figure 1B**). The amount of high-fidelity DNA polymerase was previously demonstrated to be optimal at 0.15 U in a 25 μl reaction system, at which the mismatched bases were removed from the primers, and the completion with other DNA polymerase such as Taq DNA polymerase for DNA synthesis did not occur (Hao et al., 2015; Li et al., 2019). The same amount of high-fidelity DNA polymerase was also demonstrated to be optimal for the mismatch-tolerant LAMP method (data not shown). Therefore, a standard mismatch-tolerant LAMP reaction system that includes all reagents in the conventional LAMP reaction with an additional 0.15 U high-fidelity DNA polymerase were established. Using the novel system, we developed an novel RT-LAMP assay for pan-serotype detection of DENV using the primer sets previously described (Teoh et al., 2013).

Validation of the Mismatch-Tolerant RT-LAMP Assay

To assess the influence of various mismatches between primers and templates on LAMP amplification, and validate the effectiveness of the novel mismatch-tolerant RT-LAMP, we constructed a series of mutant RNA of DENV-4 that can form three kinds of mismatches with the 3′-ends of F3, FIP, and BLP, as well as the 5′-end of FIP (i.e., 3′-end of F1) (**Figure 2A**). We firstly used the conventional RT-LAMP assay to detect 3 × 10^4 copies of the mutant RNAs, and a control wild type template. Relative to the wild type template (Tt: 21.7 min), the mutants generated slower RT-LAMP amplification with Tt values of 22.8–44.3 min (**Figure 2B** and **Table 1**). For the same primer, three kinds of mismatches formed with the mutant RNAs resulted in a relatively large variance in Tt values (0.5–6.5 min) (**Table 1**). In particular, we found that the mismatches occurring in different primers had marked difference in their inhibition effect on LAMP amplification. The mutants forming mismatches with the 3′-end of the F3 primer had Tt values of 22.8–25.2 min (about 1–3.5 min higher than the wild type), whereas the mutants forming mismatches with the 3′-end of the BLP or FIP (corresponding to the 3′-end of F2) primers had Tt values of 35.2–44.3 min, which are 13.5–22.6 min higher than the wild type (**Table 1**).

Then, we used the mismatch-tolerant RT-LAMP assay to test the same mutants and the wild type template. The novel method had Tt value of 16.1 min for the wild type template, about 5.6 min lower than that in the conventional method (**Figure 2B** and **Table 1**). A dramatic improvement on the LAMP amplification was achieved by the novel method for all mutants (about 5.9–22.6 min faster than the conventional method) ($P < 0.001$, paired T-Test) (**Table 1**). The mutants forming mismatches at the 3-ends of both the BLP and FIP, as well as at the 5′-end of the FIP (corresponding to the 3′-end of F1) had about 13.7–22.6 min

[1] http://vassarstats.net/clin1.html

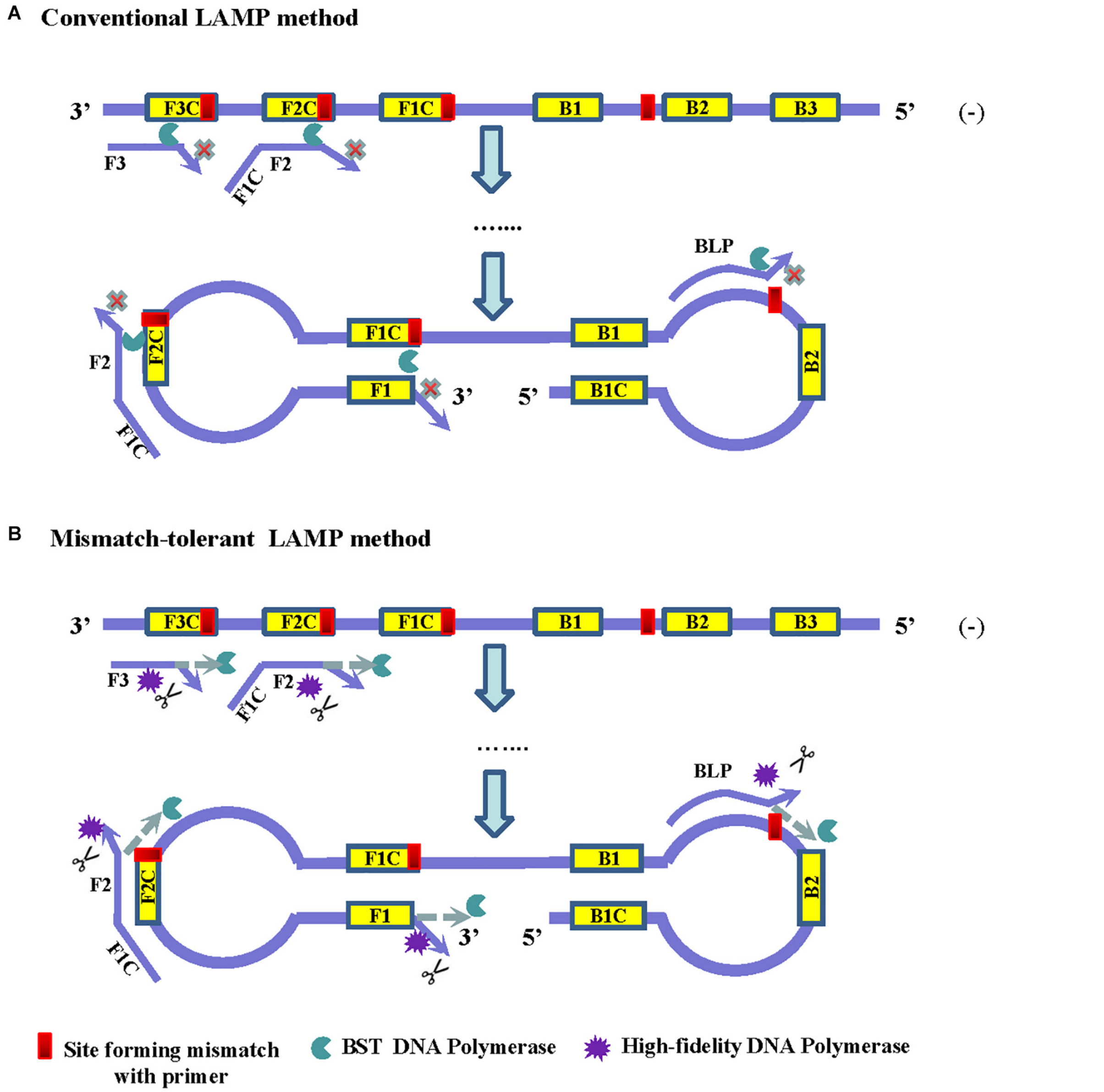


FIGURE 1 | Principles of the conventional **(A)** and the mismatch-tolerant LAMP **(B)** methods. The presence of a mismatch at 3′-end of the primer will largely reduce or stop LAMP amplification. The presence of a small amount of high-fidelity DNA polymerase, which has 3′–5′ exonuclease activity, can remove the mismatched bases from the primers, allowing the Bst DNA polymerase to initiate primer extension.

lower Tt values by the novel method than the conventional assay, and the mutants having mismatches with the F3 primer showed about 5.9–8.5 min lower Tt values than the conventional ones (**Table 1**). Importantly, the novel method seemed to have a similar amplification efficiency for the wild type and the mutants, except for slightly slower (about 4.3–6.5 min) for the mutants forming mismatches with the 3′-ends of the FIP and BLP than for the wild type (**Table 1**).

Pan-Serotype Detection of DENV

The LAMP primers for pan-serotype detection of DENV were previously designed in the 3′ UTR region of the DENV genome, the most conserved genomic region shared among the four serotypes (**Figure 3A**). To further evaluate the primers, we retrieved all available 3′ UTR sequences of four DENV serotypes, and performed sequence alignments. A total of 1220 DENV-1, 626 DENV-2, 659 DENV-3, and 130 DENV-4 sequences were obtained (**Figure 3B**). Except for the F3 region, the vast majority of sequences are conserved for all four serotypes, and only a few variants in each serotype can form mismatches with the LAMP primers (**Figure 3B**). Furthermore, DENV-2 and DENV-4 appeared to be distinct from DENV-1 and DENV-3 in the F3 and F2 regions (especially the regions corresponding to the 5′ parts of the primers) (**Figure 3B**). When all four DENV strains

TABLE 1 | The influence of various mismatches between primers and templates on the amplification times by the novel mismatch-tolerant and the conventional RT-LAMP methods.

Template	The novel mismatch-tolerant RT-LAMP						The conventional RT-LAMP						Tt Diff. (Novel-Conv.)
	Tt-T1 (min)	Tt-T2 (min)	Tt-T3 (min)	Mean	SD	Tt Diff. (Mut-WT)	Tt-T1 (min)	Tt-T2 (min)	Tt-T3 (min)	Mean	SD	Tt Diff. (Mut-WT)	
Wild	17.02	17.1	14.09	16.1	1.7	**NA**	22.1	22.9	20.0	21.7	1.5	**NA**	**−5.6**
F1-Mu-A	15.51	15.28	13.87	14.9	0.9	**−1.2**	26.5	33.8	25.7	28.6	4.5	**6.9**	**−13.7**
F1-Mu-G	17.22	16.4	14.86	16.2	1.2	**0.1**	33.5	37.1	34.7	35.1	1.8	**13.4**	**−18.9**
F1-Mu-C	16.82	17.43	14.61	16.3	1.5	**0.2**	31.8	33.5	31.3	32.2	1.1	**10.5**	**−15.9**
F2-Mu-A	21.1	21.73	18.14	20.3	1.9	**4.3**	36.4	38.9	30.3	35.2	4.4	**13.5**	**−14.9**
F2-Mu-T	20.64	22.52	18.56	20.6	2.0	**4.5**	32.8	41.1	34.3	36.1	4.4	**14.4**	**−15.5**
F2-Mu-G	20.73	20.82	19.78	20.4	0.6	**4.4**	44.9	36.7	35.3	39.0	5.2	**17.3**	**−18.5**
F3-Mu-A	17.55	15.93	15.21	16.2	1.2	**0.2**	27.9	25.3	21.0	24.7	3.5	**3.0**	**−8.5**
F3-Mu-G	16.93	17.57	16.35	17.0	0.6	**0.9**	23.0	23.7	21.7	22.8	1.0	**1.1**	**−5.9**
F3-Mu-C	18.09	16.66	15.44	16.7	1.3	**0.7**	25.4	28.1	22.1	25.2	3.0	**3.5**	**−8.4**
BLP-Mu-A	23.45	21.55	19.09	21.4	2.2	**5.3**	46.7	45.1	35.0	42.2	6.3	**20.6**	**−20.9**
BLP-Mu-T	21.81	24.04	19.48	21.8	2.3	**5.7**	46.8	49.7	36.6	44.3	6.9	**22.6**	**−22.6**
BLP-Mu-G	21.97	24.18	21.43	22.5	1.5	**6.5**	37.6	40.3	39.1	39.0	1.3	**17.3**	**−16.5**

SD, standard deviation; NA, not applicable; Diff., difference; Mut, mutants; WT, wild type; Tt, time threshold; T1-T3, replicates 1–3; Conv., conventional RT-LAMP method; Novel, the mismatch-tolerant RT-LAMP method; min, minute. The Tt differences are highlighted in bold.

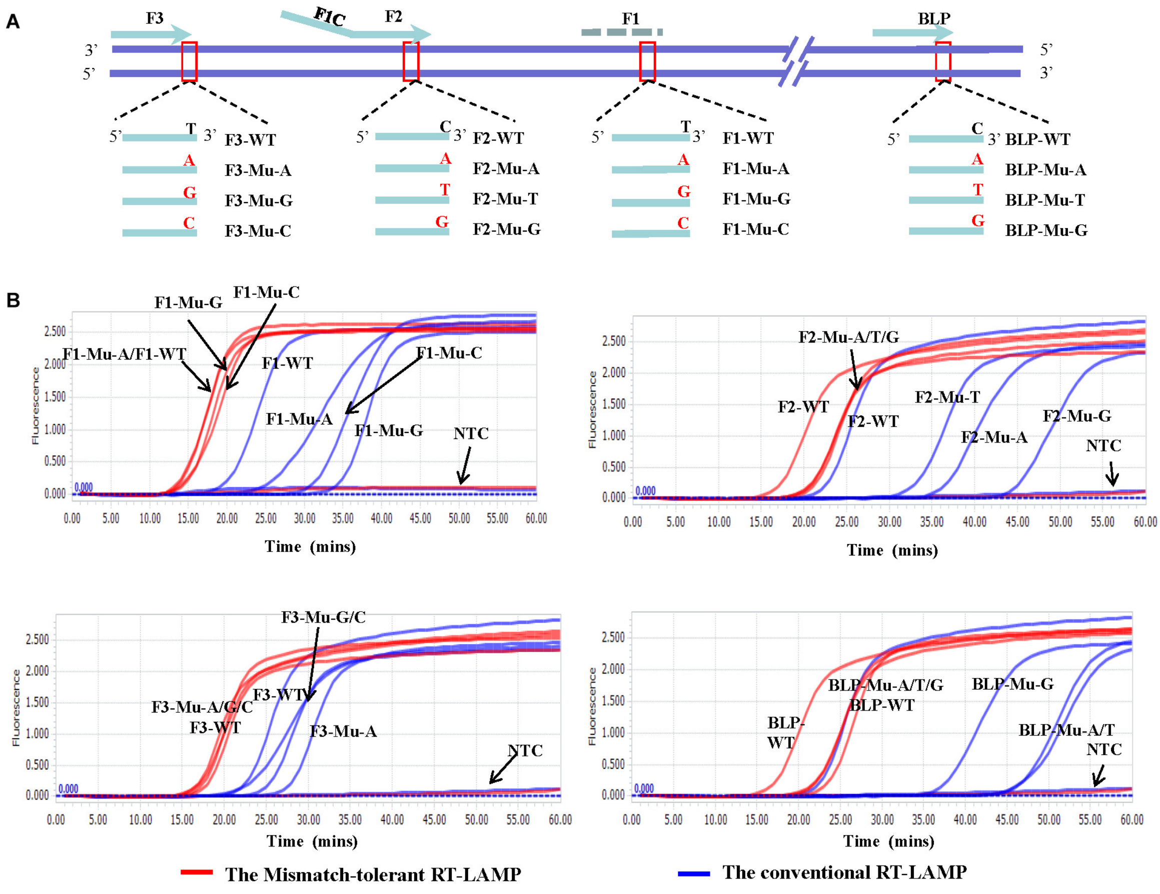


FIGURE 2 | Flexibility of the novel mismatch-tolerant RT-LAMP to various mismatches. **(A)** Primer and template sets. **(B)** Amplification curves of the novel mismatch-tolerant and the conventional RT-LAMP methods for various mutants that form mismatches with the primers. About 30,000 RNA copies of mutant or wild-type template were added in each reaction. WT, wild-type; Mu, mutant; NTC, no template control.

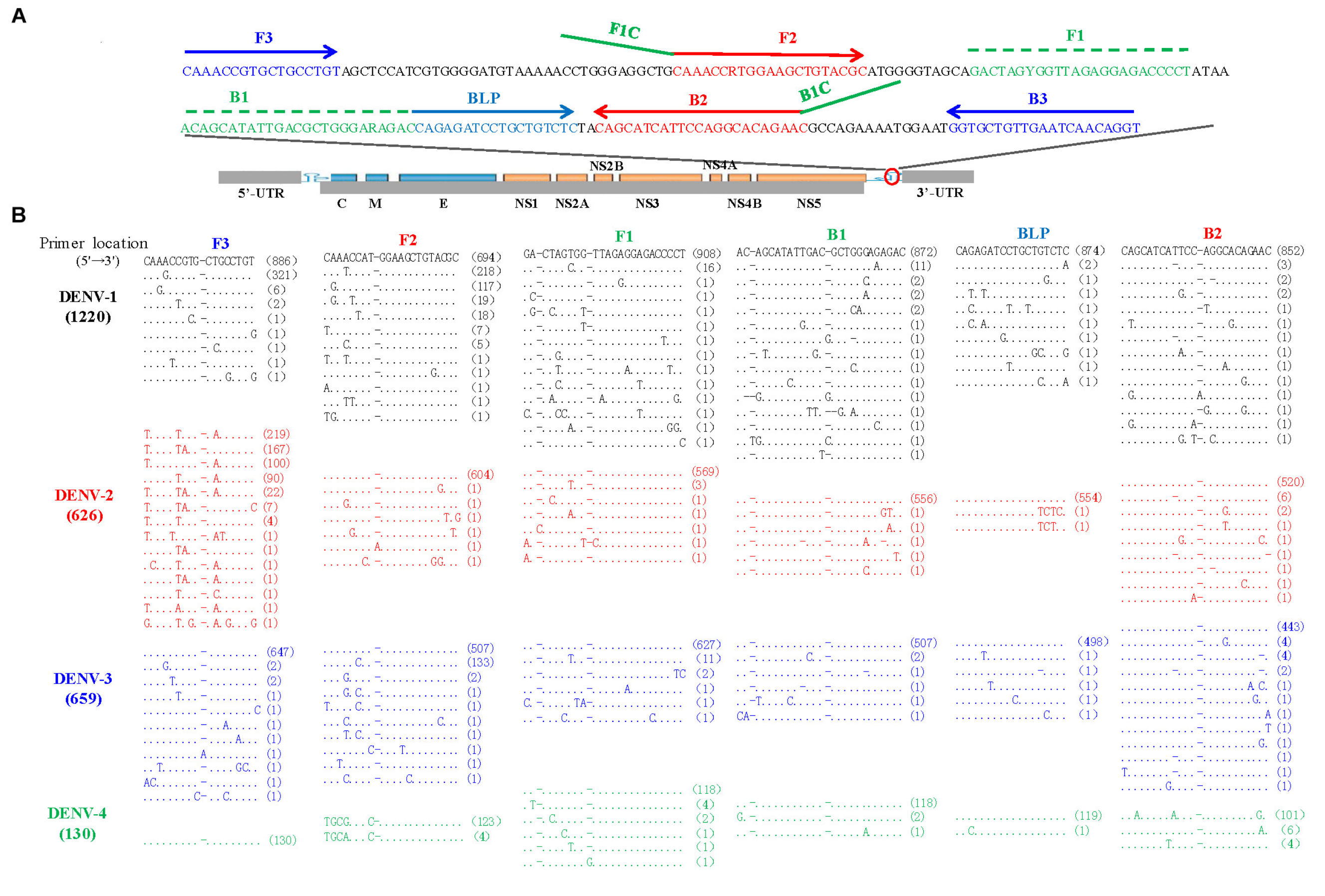

A
F3
F1C
F2
F1
CAAACCGTGCTGCCTGTAGCTCCATCGTGGGGATGTAAAAACCTGGGAGGCTGCAAACCRTGGAAGCTGTACGCATGGGGTAGCAGACTAGYGGTTAGAGGAGACCCCTATAA
B1
BLP
B2
B1C
B3
ACAGCATATTGACGCTGGGARAGACCAGAGATCCTGCTGTCTCTACAGCATCATTCCAGGCACAGAACGCCAGAAAATGGAATGGTGCTGTTGAATCAACAGGT
NS2B
NS4A
5'-UTR
3'-UTR
C
M
E
NS1
NS2A
NS3
NS4B
NS5
B
Primer location (5'→3')
F3
F2
F1
B1
BLP
B2
CAAACCGTG-CTGCCTGT (886)
CAAACCAT-GGAAGCTGTACGC (694)
GA-CTAGTGG-TTAGAGGAGACCCCT (908)
AC-AGCATATTGAC-GCTGGGAGAGAC (872)
CAGAGATCCTGCTGTCTC (874)
CAGCATCATTCC-AGGCACAGAAC (852)
DENV-1 (1220)
DENV-2 (626)
DENV-3 (659)
DENV-4 (130)

FIGURE 3 | Locations **(A)** and sequence alignments **(B)** of the primer regions of all available DENV strains. A total of 1220 DENV-1 (black), 626 DENV-2 (red), 659 DENV-3 (Blue), and 130 DENV-4 (green) sequences were downloaded from GenBank on September 10, 2018. Dot, identity with the topmost sequence; Dash, deletion.

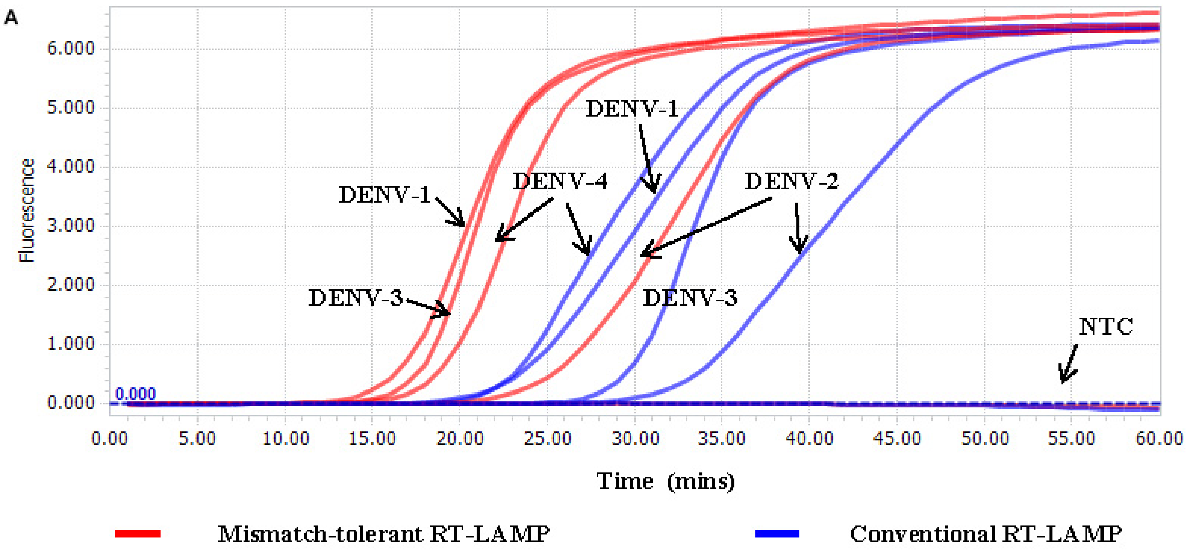
were detected using the conventional RT-LAMP assay, a lower amplification efficiency was observed for DENV-2 (**Figure 4A**). Sequencing and sequence analysis showed that there were four substitutions in the FIP and BIP regions of DENV-2 compared to other serotypes (**Supplementary Figure S1**). To examine whether the mismatch-tolerant RT-LAMP can improve the amplification efficiency for all four DENV serotypes, we used the novel method to detect the same DENV strains and compared it with the conventional RT-LAMP assay. All four serotypes had lower Tt values (14.8–23.2 min) by the novel RT-LAMP assay than that by the conventional ones (21.3–31.5 min) (**Figure 4B**). The differences in Tt values between the two assays were 4.5–11.7 min. To avoid non-specific amplification, the reaction time of the new method was set at 50 min for the subsequent experiments.

Cross-Reactivity and Limit of Detection of the Novel RT-LAMP Assay

Cross-reactivity test showed that there was no amplification of other ten human viruses (including closely related flaviviruses, Zika virus, and yellow fever virus) by the novel DENV RT-LAMP assay within 50 min (**Figure 5**), indicating that the new assay is specific for DENV. Sensitivity tests showed that the novel pan-serotype DENV RT-LAMP can detect as low as 162, 886, 78, and 104 RNA copies of DENV-1, DENV-2, DENV-3, and DENV-4, which are equivalent to 3.36, 84, 0.17, and 1.34 PFU, respectively. In comparison, the conventional RT-LAMP assay only yielded one or two positive amplifications among the three replicates at the same concentrations of DENV-1, DENV-3 and DENV-4, except for DENV-2, for which 886 virus RNA copies were detected by all replicates (**Figure 6**). Of particular importance is that all the amplification curves by the novel RT-LAMP assay appeared within 25 min, whereas almost all curves by the conventional assay appeared after 25 min (**Figure 6**). The LOD values of the novel assay, as described in the materials and methods, were determined as 74, 252, 78, and 35 RNA copies per reaction for DENV-1, DENV-2, DENV-3, and DENV-4, respectively (**Table 2**).

Visual Detection

To prepare for use in the resource-poor settings, the novel RT-LAMP assay was further developed into a visual determination assay by adding HNB or cresol red. Because both HNB and Cresol Red assays gave a clear color indication for all samples at 50 min (**Figure 7** and **Supplementary Figure S2**), it was



B

Serotypes	The novel mismatch-tolerant RT-LAMP	The conventional RT-LAMP	Tt Diff. (Novel-conv.)
DENV-1	14.79	21.33	-6.54
DENV-2	23.20	31.51	-8.31
DENV-3	16.33	28.04	-11.71
DENV-4	17.19	21.70	-4.51

FIGURE 4 | Detection of four DENV serotypes using the novel mismatch-tolerant and the conventional RT-LAMP assays. **(A)** Amplification curves of the novel and the conventional RT-LAMP methods. **(B)** Comparison of the Tt values between the novel and the conventional RT-LAMP. About 3,000 RNA copies of each DENV serotype was used in the test. Mins, minutes; NTC, no template control.

now selected as the cut-off for the visual assays. DENV-1, DENV-2, DENV-3, and DENV-4 were detected by the novel assay at concentrations of 162, 89, 78, and 104 copies, respectively (**Figure 7**), showing the same sensitivity as the novel RT-LAMP with real-time monitoring. As a comparison, the conventional assay only detected 1620, 886, 780, and 1040 copies of DENV-1, DENV-2, DENV-3, and DENV-4, respectively, for the same duration of amplification.

TABLE 2 | Limit of detection (LOD) of the novel mismatch-tolerant RT-LAMP.

Dilution	Standard (copies/reaction)	Positive/total tested			
		DENV-1	DENV-2	DENV-3	DENV-4
	30,000	10/10	10/10	10/10	10/10
5×	6000	10/10	10/10	10/10	10/10
5×	1200	10/10	10/10	10/10	10/10
5×	240	10/10	9/10	10/10	10/10
5×	48	4/10	0/10	8/10	10/10
5×	9.6	0/10	0/10	4/10	1/10
LOD (copies/reaction)		74	252	78	35

Evaluation of the Novel Pan-Serotype DENV RT-LAMP Assay

To explore the potential for clinical application of our novel assay, a total of 153 plasma samples collected from dengue-suspected patients were used to evaluate the novel DENV RT-LAMP assay. Of these plasma samples, 106 (69.3%) were detected as DENV positive by all four assays (**Supplementary Figure S3**), and various proportions were detected as being positive by individual assay.

The novel RT-LAMP, the NS1 antigen detection, the specific RT-PCR and the conventional RT-LAMP assays detected 145 (94.8%), 141 (92.2%), 120 (78.4%), and 133 (86.9%) DENV positive samples, respectively (**Table 3**). The concordance rates of the novel RT-LAMP assay were 90.8% [95% confidence interval (CI): 84.8–94.7%; kappa value: 0.253] and 78.4% (95% CI: 70.9–84.5%; kappa value: 0.121) with the NS1 antigen detection and the specific RT-PCR assays, respectively (**Table 3**). The positivity results were more consistent between the novel and the conventional RT-LAMP assays, with a concordance rate of 92.2% (95% CI: 86.4–95.7%; kappa value: 0.537).

To assess the capacity of the novel RT-LAMP assay for detection of all four DENV serotypes, 120 samples that were previously genotyped were used, and 116 of them (96.7%) were also tested as positive by the novel RT-LAMP assay (**Table 4**). Among them, all DENV-2 (5/5, 100%) and DENV-3 (46/46, 100%), most DENV-1 (55/58, 94.8%), and DENV-4 (10/11, 90.9%) samples were detected as being positive (**Table 4**). The novel RT-LAMP assay showed better performance for DENV-1 (94.8% vs. 87.9%) and DENV-3 (100% vs. 87.0%) than the conventional assay, and identical performance for DENV-2 and DENV-4 (**Table 4**). In addition, we randomly selected 12 samples with a negative RT-PCR test but positive by the novel RT-LAMP assay for further amplification using a RT-nested PCR assay. One additional sample was detected as being positive and further genotyped as DENV-4 by Sanger sequencing and Blast analysis. Among seven samples with a Tt difference of more than 15 min between the novel and the conventional RT-LAMP assays, three DENV-1 and four DENV-3 strains were identified, and only the three DENV-1 strains carried A- > C or C- > T substitution in the F2 region (**Supplementary Figure S4**). The two substitutions were also found in some previously sequenced DENV-1 variants (**Figure 3**).

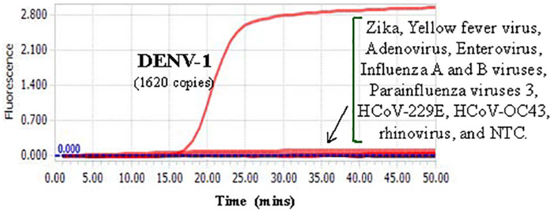


FIGURE 5 | Cross-reactivity test of the novel DENV RT-LAMP assay to other human viruses. About 1620 RNA copies (33.6 PFU) of DENV-1 was used as the positive control. Testing viruses included Zika, Yellow fever virus, Adenovirus, Enterovirus, Influenza A and B viruses, Parainfluenza viruses 3, HCoV-229E, HCoV-OC43, and rhinovirus. NTC, no template control.

TABLE 3 | Comparison among different DENV detection assays for 153 clinical samples.

Methods		NS1 antigen assay			RT-PCR			The conventional RT-LAMP			Total	Positive rate (%)
	Items	Pos.	Neg.	Concordance rate (%)	Pos.	Neg.	Concordance rate (%)	Pos.	Neg.	Concordance rate (%)		
The novel mismatch-tolerant RT-LAMP	Pos.	136	9	90.8%	116	29	78.4%	133	12	92.2%	145	94.8%
	Neg.	5	3		4	4		0	8		8	
Total		141	12	NA	120	33	NA	133	20	NA	153	NA
Positive rate (%)		92.2%		NA	78.4%		NA	86.9%		NA	NA	NA

The concordance rate was calculated using the formula: (number of consistent results by both methods/total number) × 100%. Please also see the Venn diagram in ***Supplementary Figure S3*** *for details. Pos., positive; Neg., negative; NA, not applicable.*

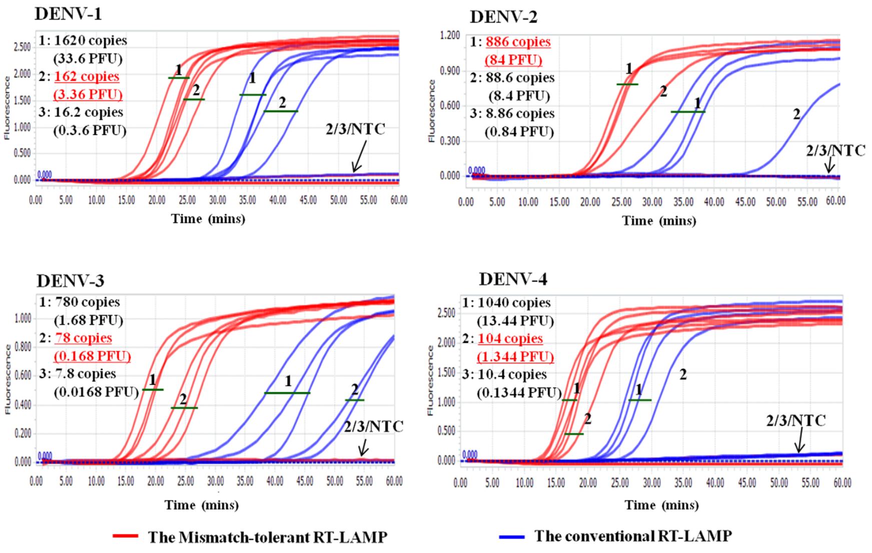


FIGURE 6 | Sensitivity comparison between the novel mismatch-tolerant and the conventional RT-LAMP assays for four DENV serotypes. The sensitivity of the novel mismatch-tolerant RT-LAMP assay for each DENV serotype is highlighted in red and underline. Positive amplification was defined only when all three replicates are successfully amplified. NTC: no template control.

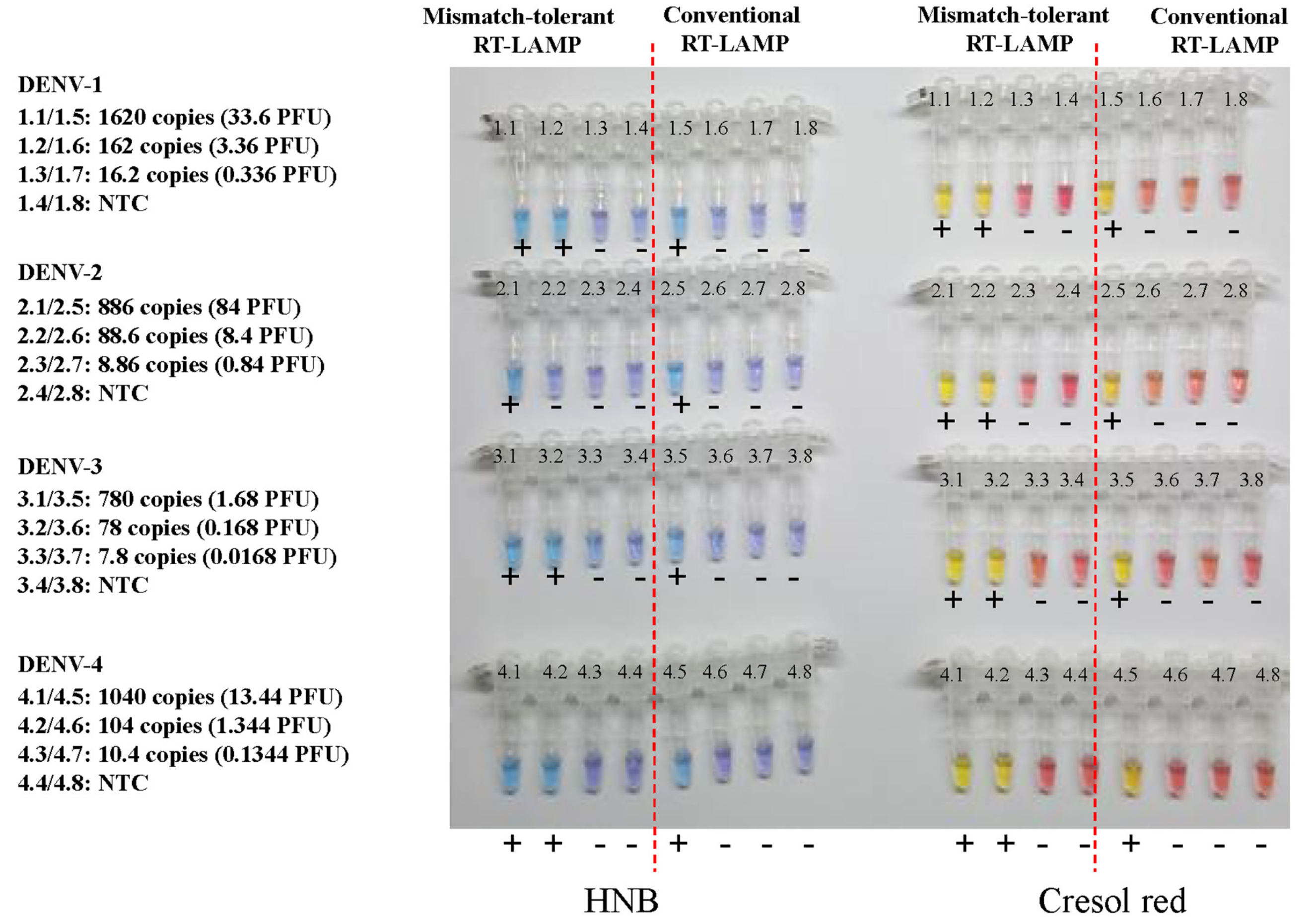

Mismatch-tolerant RT-LAMP
Conventional RT-LAMP
Mismatch-tolerant RT-LAMP
Conventional RT-LAMP
DENV-1
1.1/1.5: 1620 copies (33.6 PFU)
1.2/1.6: 162 copies (3.36 PFU)
1.3/1.7: 16.2 copies (0.336 PFU)
1.4/1.8: NTC
DENV-2
2.1/2.5: 886 copies (84 PFU)
2.2/2.6: 88.6 copies (8.4 PFU)
2.3/2.7: 8.86 copies (0.84 PFU)
2.4/2.8: NTC
DENV-3
3.1/3.5: 780 copies (1.68 PFU)
3.2/3.6: 78 copies (0.168 PFU)
3.3/3.7: 7.8 copies (0.0168 PFU)
3.4/3.8: NTC
DENV-4
4.1/4.5: 1040 copies (13.44 PFU)
4.2/4.6: 104 copies (1.344 PFU)
4.3/4.7: 10.4 copies (0.1344 PFU)
4.4/4.8: NTC
1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8
+ + - - + - - -
2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8
+ - - - + - - -
+ + - - + - - -
3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8
+ + - - + - - -
4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8
+ + - - + - - -
HNB
Cresol red

FIGURE 7 | Visual detection of four DENV serotypes using the novel mismatch-tolerant and the conventional RT-LAMP assays with HNB or cresol red. The reactions were performed at 63° for 50 min. The color changes from violet to azure for HNB and from burgundy to orange or yellow for cresol red were considered as positive (+). The results at 30, 40, and 60 min are shown in **Supplementary Figure S2**.

TABLE 4 | Detection rates of different DENV serotypes by the novel mismatch-tolerant and the conventional RT-LAMP assays.

DENV serotypes	The novel mismatch-tolerant RT-LAMP			The conventional RT-LAMP			Total
	Positive	Negative	Detection rate (%)	Positive	Negative	Detection rate (%)	
DENV-1	55	3	94.8	51	7	87.9	58
DENV-2	5	0	100	5	0	100	5
DENV-3	46	0	100	40	6	87.0	46
DENV-4	10	1	90.9	10	1	90.9	11
Total	116	4	96.7	106	14	88.3	120

DISCUSSION

Emerging and re-emerging infectious diseases are major burden to global public health systems (Mehand et al., 2018). Many of them are vector-borne infectious diseases that affect a large population in the tropical and subtropical regions (e.g., Africa, Latin America, and Southeast Asia), and cause higher morbidity and mortality in low-income countries than developed countries (Savic et al., 2014). Several recent outbreaks of vector-borne infectious diseases are caused by viruses, such as DENV, Zika virus, and yellow fever virus (Weaver et al., 2018). Development of point-of-care tests (POCT) for these viruses plays a crucial role in the prevention and control of vector-borne infectious diseases in resource-poor countries/regions (Kozel and Burnham-Marusich, 2017). In this study, we developed a mismatch-tolerant LAMP method and demonstrated its practical use by improving the sensitivity and specificity of a previous established RT-LAMP assay for the detection of all four DENV serotypes.

Our novel assay has notable features that are superior to the conventional LAMP method, which was developed in about two decades ago (Notomi et al., 2000). The conventional LAMP assay has been widely used in various fields because of its noted advantages including relatively independent of expensive instruments, stable reaction system, and flexibility in real-time monitoring of reactions using fluorescent dyes (e.g., SYTO 9 and SYBR green I) or visual determination using colorimetric dyes (e.g., HNB, Calcein or pH-sensitive dyes) (Tomita et al., 2008; Fischbach et al., 2015; Tanner et al., 2015). LAMP requires three pairs of primers: long FIP and BIP primers that bind to a relatively short genomic region (usually less than 300 bp); the 3′-ends of the F3/B3, FIP/BIP and FLP/BLP that provide 3′-hydroxyl groups and start DNA extension; the 5′-ends of the primers FIP and BIP (i.e., 3′-end of F1 and B1) that are responsible for the self-priming of the dumb-bell form DNA generated in the first stage of LAMP reaction (Notomi et al., 2000). However, the presence of mismatches at these ends markedly reduced the amplification efficiency of LAMP, as we have observed in this study and other have reported previously. In particular, the mismatches occurring at both ends of FIP/BIP and the 3′-ends of FLP/BLP had larger inhibition effect on LAMP amplification than those occurring at 3′-ends of F3/B3. One plausible reason for this phenomenon is that F3/B3 are only responsible for the initiation of LAMP cycling by strand replacement of newly synthesized DNA along the target, but do not participate in the self-priming of the dumb-bell form DNA (Notomi et al., 2000). Therefore, the conventional LAMP is very susceptible to mismatches between primers and templates.

High susceptibility of LAMP to mismatches limits its power in diagnostic application for viral infectious diseases (de Paz et al., 2014), and makes it especially difficult to apply on highly variable viral genome, for which the design of universal primers for a single assay that covers all genotypes, subtypes or rare variants of the same virus is challenging. Moreover, many viruses exist as quasispecies in which various variants can form mismatches with LAMP primers that designed with even the most conserved genomic region. In the last decade, a strategy by combining multiple degenerate primers together was developed for broad-spectrum detection of various genotypes or subtypes of genetically diverse viruses such as DENV, HIV-1, influenza A viruses, and enteroviruses (Poon et al., 2005; Teoh et al., 2013; Yamazaki et al., 2013; Dauner et al., 2015; Ocwieja et al., 2015; Curtis et al., 2018). However, low detection efficiency of LAMP for various viral variants still remains to be solved.

We have previously reported that by removing the mismatched bases in the 3′-end of a primer using high-fidelity DNA polymerase that has a 3′–5′ exonuclease activity, the performance of PCR for detection of genetically diverse viruses can be significantly improved (Li et al., 2019). A similar idea had been applied by others in the proof-reading PCR and the high-fidelity DNA polymerase-mediated qPCR (Bi and Stambrook, 1998; Zhang et al., 2017). In the current study, to improve the performance of LAMP in detection of genetically diverse viruses, we used similar principle and developed a mismatch-tolerant RT-LAMP method. The most distinctive feature of the novel method in comparison to the conventional method is that a minuscule amount of high-fidelity DNA polymerase was added into the standard LAMP system. In fact, only 0.15 U of high-fidelity DNA polymerase per 25 μl LAMP reaction mix is sufficient to accomplish optimal performance. As expected, the addition of high-fidelity DNA polymerase largely improved LAMP amplification efficiency for variable templates that form mismatches with the primers, thus demonstrating an excellent tolerance for various mismatches between primers and templates. Furthermore, compared to the conventional LAMP method, the mismatch-tolerant method not only significantly improved the detection sensitivity, but also markedly shortened the reaction time regardless of the presence or absence of mismatches between the primers and templates. The underlying mechanisms for an improved amplification for wild-type template have not been determined. Nevertheless, our findings indicated that the novel mismatch-tolerant RT-LAMP method is faster and more efficient than the conventional methods.

Having completed the proof-of-concept studies, we went on to test the robustness of the novel assay on clinical samples by choosing a genetically diverse virus, DENV, which causes an infection of great impact to public health in tropical and subtropical area (Bhatt et al., 2013). Early and rapid detection of DENV infection is important for patient management and epidemiological monitoring, especially in resource-limited settings (Shu and Huang, 2004; Teles et al., 2005). Current assays for dengue diagnosis have their limitations. NS1 antigen detection assay is highly sensitive and specific for acute DENV infection, but the pre-existing anti-NS1 IgG antibody induced by a previous infection will interfere with the detection of subsequent DENV infection especially in high-prevalence regions where many DENV serotypes co-circulate (Teles et al., 2005; Fuchs et al., 2014). In principle, NAA tests are more sensitive and specific for early and accurate detection of DENV infection than the NS1 antigen assay. Therefore, we used the novel method to improve a previously established RT-LAMP assay for pan-serotype detection of DENV (Teoh et al., 2013). Compared to the previous assay, the novel assay showed a significant improvement for detection of all four DENV serotypes. DENV infected patients often had mean viremia of $10^{7.6-8.5}$ PFU per ml of plasma at the acute phase (Vaughn et al., 2000). LOD tests revealed that the novel assay is sufficiently sensitive for detecting these viremia levels. Moreover, the evaluation of 153 clinical samples showed that the novel assay had higher detection rate than the NS1 antigen assay and RT-PCR method for all four DENV serotypes.

In resource-limited settings, diagnosis of dengue relies mainly on the clinical manifestation, which is easily confused with similar symptoms caused by other infectious agents, such as malaria, chikungunya, and influenza (Henchal and Putnak, 1990; Gubler, 1998). The novel RT-LAMP method has a high sensitivity of detection for multiple viral genotype/subtypes with a short amplification time, and can be developed into a visual measurement qualitative assay by adding HNB, cresol red, or other pH-sensitive dyes (e.g., phenol red and neutral red). The above features offer great application potential for our assay in POCT at local clinics and even in the fields. In particular, recent progresses in the nucleic acid extraction-free protocols and lyophilized formulation of LAMP reagents will further facilitate its application in the diagnosis of various viral infectious diseases in different settings in resource-poor countries (Carter et al., 2017; Liu et al., 2018).

Despite having many significant findings, this study also has several limitations. First, because there lacks a gold standard to determine the true positive and true negative samples for DENV infection, we were unable to calculate the sensitivity, specificity and accuracy of the novel DENV RT-LAMP assay. Second, the sample size used in the clinical evaluation was relatively small, especially DENV-2 samples and DENV-negative samples. Further validation of our assay with a large sample size is needed. Third, in this study, we only applied the novel RT-LAMP method to DENV detection. Its application to the detection of other genetically diverse viruses may help to further evaluate the robustness of our new assay.

CONCLUSION

A novel mismatch-tolerant LAMP was developed by simply adding a minuscule amount of high-fidelity DNA polymerase in addition to the standard amount of Bst DNA polymerase to the standard LAMP reaction mixture. This method could be applied to update all LAMP-based diagnostic assays. Importantly, the novel LAMP method tolerates well mismatches between primers and templates, and has higher amplification efficiency for viral variants than the conventional LAMP method. Therefore, the mismatch-tolerant LAMP method represents a simple, sensitive and promising approach for molecular diagnosis of genetically diverse viruses, and it is especially suited for application in resource-limited settings.

ETHICS STATEMENT

This study was approved by the Ethics Committees of Shanghai Public Health Clinical Center and Kunming University of Science and Technology. Written informed consents were obtained from all dengue-suspected patients.

AUTHOR CONTRIBUTIONS

CZ conceived and designed the study. YZ, SY, ML, and BW performed the experiments. CZ, YZ, ZW, YL, YH, XJ, and NY analyzed and interpreted the data. CZ, XJ, and XX contributed reagents and materials. CZ and ZW drafted the manuscript. XJ contributed to critical revision of the manuscript. CZ and NY supervised the study.

FUNDING

This work was supported by the grants from the National Science and Technology Major Project of China (2018ZX10101004003001 and 2017ZX10103009-002), the “One Belt One Road” Project (153831KYSB20170043) of the Chinese Academy of Sciences, and the 133 project of Institut Pasteur of Shanghai, CAS. 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 Mengling Zhang, Kai Liu, Yanpeng Li, and Xuemin Fu at Institut Pasteur of Shanghai, CAS for their technique help and valuable discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01056/full#supplementary-material

REFERENCES

Anderson, D. J. (1989). Determination of the lower limit of detection. *Clin. Chem.* 35, 2152–2153.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., et al. (2013). The global distribution and burden of dengue. *Nature* 496, 504–507. doi: 10.1038/nature12060

Bi, W., and Stambrook, P. J. (1998). Detection of known mutation by proof-reading PCR. *Nucleic Acids Res.* 26, 3073–3075. doi: 10.1093/nar/26.12.3073

Bustin, S. A. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* 25, 169–193. doi: 10.1677/jme.0.0250169

Bustin, S. A. (2002). Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J. Mol. Endocrinol.* 29, 23–39. doi: 10.1677/jme.0.0290023

Carter, C., Akrami, K., Hall, D., Smith, D., and Aronoff-Spencer, E. (2017). Lyophilized visually readable loop-mediated isothermal reverse transcriptase nucleic acid amplification test for detection Ebola Zaire RNA. *J. Virol. Methods* 244, 32–38. doi: 10.1016/j.jviromet.2017.02.013

Curtis, K. A., Morrison, D., Rudolph, D. L., Shankar, A., Bloomfield, L. S. P., Switzer, W. M., et al. (2018). A multiplexed RT-LAMP assay for detection of group M HIV-1 in plasma or whole blood. *J. Virol. Methods* 255, 91–97. doi: 10.1016/j.jviromet.2018.02.012

Dauner, A. L., Mitra, I., Gilliland, T Jr, Seales, S., Pal, S., Yang, S. C., et al. (2015). Development of a pan-serotype reverse transcription loop-mediated isothermal amplification assay for the detection of dengue virus. *Diagn. Microbiol. Infect. Dis.* 83, 30–36. doi: 10.1016/j.diagmicrobio.2015.05.004

de Paz, H. D., Brotons, P., and Munoz-Almagro, C. (2014). Molecular isothermal techniques for combating infectious diseases: towards low-cost point-of-care diagnostics. *Expert. Rev. Mol. Diagn.* 14, 827–843. doi: 10.1586/14737159.2014.940319

Domingo, E., and Perales, C. (2018). Quasispecies and virus. *Eur. Biophys. J.* 47, 443–457. doi: 10.1007/s00249-018-1282-6

Fischbach, J., Xander, N. C., Frohme, M., and Glokler, J. F. (2015). Shining a light on LAMP assays–a comparison of LAMP visualization methods including the novel use of berberine. *Biotechniques* 58, 189–194. doi: 10.2144/000114275

Fuchs, I., Bin, H., Schlezinger, S., and Schwartz, E. (2014). NS1 antigen testing for the diagnosis of dengue in returned Israeli travelers. *J. Med. Virol.* 86, 2005–2010. doi: 10.1002/jmv.23879

Go, Y. Y., Rajapakse, R., Kularatne, S. A. M., Lee, P. A., Ku, K. B., Nam, S., et al. (2016). A pan-dengue virus reverse transcription-insulated isothermal PCR assay intended for point-of-need diagnosis of dengue virus infection by use of the POCKIT nucleic acid analyzer. *J. Clin. Microbiol.* 54, 1528–1535. doi: 10.1128/JCM.00225-16

Gubler, D. J. (1998). Dengue and dengue hemorrhagic fever. *Clin. Microbiol. Rev.* 11, 480–496.

Hao, W., Fan, L., Chen, Q., Chen, X., Zhang, S., Lan, K., et al. (2015). Modified proofreading PCR for detection of point mutations, insertions and deletions using a ddNTP-Blocked Primer. *PLoS One* 10:e0123468. doi: 10.1371/journal.pone.0123468

Henchal, E. A., and Putnak, J. R. (1990). The dengue viruses. *Clin. Microbiol. Rev.* 3, 376–396.

Kozel, T. R., and Burnham-Marusich, A. R. (2017). Point-of-care testing for infectious diseases: past, present, and future. *J. Clin. Microbiol.* 55, 2313–2320. doi: 10.1128/JCM.00476-17

Kralik, P., and Ricchi, M. (2017). A basic guide to real time pcr in microbial diagnostics: definitions, parameters, and everything. *Front. Microbiol.* 8:108. doi: 10.3389/fmicb.2017.00108

Li, Y., Wan, Z., Hu, Y., Zhou, Y., Chen, Q., and Zhang, C. (2019). A mismatch-tolerant RT-quantitative PCR: application to broad-spectrum detection of respiratory syncytial virus. *Biotechniques* 66, 225–230.

Ligon, B. L. (2005). Dengue fever and dengue hemorrhagic fever: a review of the history, transmission, treatment, and prevention. *Semin. Pediatr. Infect. Dis.* 16, 60–65. doi: 10.1053/j.spid.2004.09.013

Liu, X., Zhang, C., Zhao, M., Liu, K., Li, H., Li, N., et al. (2018). A direct isothermal amplification system adapted for rapid SNP genotyping of multifarious sample types. *Biosens. Bioelectron* 115, 70–76. doi: 10.1016/j.bios.2018.05.021

Mehand, M. S., Al-Shorbaji, F., Millett, P., and Murgue, B. (2018). The WHO R&D blueprint: 2018 review of emerging infectious diseases requiring urgent research and development efforts. *Antiviral Res.* 159, 63–67. doi: 10.1016/j.antiviral.2018.09.009

Nagamine, K., Hase, T., and Notomi, T. (2002). Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Mol. Cell Probes* 16, 223–229. doi: 10.1006/mcpr.2002.0415

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N., et al. (2000). Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res.* 28:E63.

Ocwieja, K. E., Sherrill-Mix, S., Liu, C., Song, J., Bau, H., and Bushman, F. D. (2015). A reverse transcription loop-mediated isothermal amplification assay optimized to detect multiple HIV subtypes. *PLoS One* 10:e0117852. doi: 10.1371/journal.pone.0117852

Poon, L. L., Leung, C. S., Chan, K. H., Lee, J. H., Yuen, K. Y., Guan, Y., et al. (2005). Detection of human influenza A viruses by loop-mediated isothermal amplification. *J. Clin. Microbiol.* 43, 427–430. doi: 10.1128/JCM.43.1.427-430.2005

Saijo, M., Morikawa, S., and Kurane, I. (2008). Real-time quantitative polymerase chain reaction for virus infection diagnostics. *Expert. Opin. Med. Diagn.* 2, 1155–1171. doi: 10.1517/17530059.2.10.1155

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *J. Virol.* 84, 9733–9748. doi: 10.1128/JVI.00694-10

Savic, S., Vidic, B., Grgic, Z., Potkonjak, A., and Spasojevic, L. (2014). Emerging vector-borne diseases - incidence through vectors. *Front. Public Health* 2:267. doi: 10.3389/fpubh.2014.00267

Shu, P. Y., and Huang, J. H. (2004). Current advances in dengue diagnosis. *Clin. Diagn. Lab. Immunol.* 11, 642–650. doi: 10.1128/CDLI.11.4.642-650.2004

Smith, C. J., and Osborn, A. M. (2009). Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol. Ecol.* 67, 6–20. doi: 10.1111/j.1574-6941.2008.00629.x

Tanner, N. A., Zhang, Y., and Evans, T. C. Jr. (2015). Visual detection of isothermal nucleic acid amplification using pH-sensitive dyes. *Biotechniques* 58, 59–68. doi: 10.2144/000114253

Teles, F. R., Prazeres, D. M., and Lima-Filho, J. L. (2005). Trends in dengue diagnosis. *Rev. Med. Virol.* 15, 287–302. doi: 10.1002/rmv.461

Teoh, B. T., Sam, S. S., Tan, K. K., Johari, J., Danlami, M. B., Hooi, P. S., et al. (2013). Detection of dengue viruses using reverse transcription-loop-mediated isothermal amplification. *BMC Infect. Dis.* 13:387. doi: 10.1186/1471-2334-13-387

Tomita, N., Mori, Y., Kanda, H., and Notomi, T. (2008). Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nat. Protoc.* 3, 877–882. doi: 10.1038/nprot.2008.57

Vaughn, D. W., Green, S., Kalayanarooj, S., Innis, B. L., Nimmannitya, S., Suntayakorn, S., et al. (2000). Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. *J. Infect. Dis.* 181, 2–9. doi: 10.1086/315215

Wang, B., Liang, Y., Yang, S., Du, Y., Xiong, L. N., Zhao, T., et al. (2018). Co-circulation of 4 dengue virus serotypes among travelers entering China from Myanmar, 2017. *Emerg. Infect. Dis.* 24, 1756–1758. doi: 10.3201/eid2409.180252

Wang, B., Yang, H., Feng, Y., Zhou, H., Dai, J., Hu, Y., et al. (2016). The distinct distribution and phylogenetic characteristics of dengue virus serotypes/genotypes during the 2013 outbreak in Yunnan, China: phylogenetic characteristics of 2013 dengue outbreak in Yunnan, China. *Infect. Genet. Evol.* 37, 1–7. doi: 10.1016/j.meegid.2015.10.022

Weaver, S. C., Charlier, C., Vasilakis, N., and Lecuit, M. (2018). Zika, chikungunya, and other emerging vector-borne viral diseases. *Annu. Rev. Med.* 69, 395–408. doi: 10.1146/annurev-med-050715-105122

Wong, Y. P., Othman, S., Lau, Y. L., Radu, S., and Chee, H. Y. (2018). Loop-mediated isothermal amplification (LAMP): a versatile technique for detection of micro-organisms. *J. Appl. Microbiol.* 124, 626–643. doi: 10.1111/jam.13647

Yamazaki, W., Mioulet, V., Murray, L., Madi, M., Haga, T., Misawa, N., et al. (2013). Development and evaluation of multiplex RT-LAMP assays for rapid and sensitive detection of foot-and-mouth disease virus. *J. Virol. Methods* 192, 18–24. doi: 10.1016/j.jviromet.2013.03.018

Yan, L., Zhou, J., Zheng, Y., Gamson, A. S., Roembke, B. T., Nakayama, S., et al. (2014). Isothermal amplified detection of DNA and RNA. *Mol. Biosyst.* 10, 970–1003. doi: 10.1039/c3mb70304e

Yang, S., and Rothman, R. E. (2004). PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect. Dis.* 4, 337–348. doi: 10.1016/S1473-3099(04)01044-8

Zhang, M., Liu, K., Hu, Y., Lin, Y., Li, Y., Zhong, P., et al. (2017). A novel quantitative PCR mediated by high-fidelity DNA polymerase. *Sci. Rep.* 7:10365. doi: 10.1038/s41598-017-10782-4

Zhao, Y., Chen, F., Li, Q., Wang, L., and Fan, C. (2015). Isothermal amplification of nucleic acids. *Chem. Rev.* 115, 12491–12545. doi: 10.1021/acs.chemrev.5b00428

Conflict of Interest Statement: 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.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 May 2019
doi: 10.3389/fmicb.2019.01139

Assessing the Diversity of Endogenous Viruses Throughout Ant Genomes

Peter J. Flynn[1,2]* *and Corrie S. Moreau*[2,3]

[1] Committee on Evolutionary Biology, The University of Chicago, Chicago, IL, United States, [2] Department of Science and Education, Integrative Research Center, Field Museum of Natural History, Chicago, IL, United States, [3] Departments of Entomology and Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, United States

Endogenous viral elements (EVEs) can play a significant role in the evolution of their hosts and have been identified in animals, plants, and fungi. Additionally, EVEs potentially provide an important snapshot of the evolutionary frequency of viral infection. The purpose of this study is to take a comparative host-centered approach to EVE discovery in ant genomes to better understand the relationship of EVEs to their ant hosts. Using a comprehensive bioinformatic pipeline, we screened all nineteen published ant genomes for EVEs. Once the EVEs were identified, we assessed their phylogenetic relationships to other closely related exogenous viruses. A diverse group of EVEs were discovered in all screened ant host genomes and in many cases are similar to previously identified exogenous viruses. EVEs similar to ssRNA viral proteins are the most common viral lineage throughout the ant hosts, which is potentially due to more chronic infection or more effective endogenization of certain ssRNA viruses in ants. In addition, both EVEs similar to viral glycoproteins and retrovirus-derived proteins are also abundant throughout ant genomes, suggesting their tendency to endogenize. Several of these newly discovered EVEs are found to be potentially functional within the genome. The discovery and analysis of EVEs is essential in beginning to understand viral–ant interactions over evolutionary time.

Keywords: Formicidae, endogenous viral elements, comparative genome biology, microbes, viral diversity

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana–Champaign, United States

Reviewed by:
David Richard Nash,
University of Copenhagen, Denmark
Frank O'Neill Aylward,
Virginia Tech, United States
Lumi Viljakainen,
University of Oulu, Finland

****Correspondence:***
Peter J. Flynn
pflynn@uchicago.edu

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

Received: *31 January 2019*
Accepted: *06 May 2019*
Published: *22 May 2019*

Citation:
Flynn PJ and Moreau CS (2019)
Assessing the Diversity
of Endogenous Viruses Throughout
Ant Genomes.
Front. Microbiol. 10:1139.
doi: 10.3389/fmicb.2019.01139

INTRODUCTION

Endogenous viral elements (EVEs), or viral fossils, are whole or fragmented viral sequences integrated into host genomes after viral infection, which can then propagate through the germline. The majority of research conducted on endogenous viruses centers around retroviruses, which has led to discoveries demonstrating that these viruses could play a role in the evolution of their hosts. EVEs were found to be potentially important in the evolution of placental mammals as well as in the resistance to a variety of diseases (Feschotte and Gilbert, 2012; Grasis, 2017).

Endogenous viral elements are created when a duplicate of a double-stranded DNA viral genome is incorporated into the host germline. As part of their replication, retroviruses must manufacture dsDNA intermediates in order to integrate into the host genome. However, DNA and RNA viral endogenization is much less well understood. It is thought to result from inadvertent chromosomal integration such as non-homologous recombination or retrotransposition events (Aiewsakun and Katzourakis, 2015; Figure 1 from Katzourakis and Gifford, 2010). DNA repair machinery from the host cell has the ability to detect viral sequences within the genome (Weitzman et al., 2004). Therefore, EVEs will often be excised from the host genome, though a small number will evade detection. EVEs reach genomic fixation either from neutral evolution or from exaptation, a process whereby EVEs convey beneficial functions distinct from their original purpose to their host (Katzourakis and Gifford, 2010). EVEs will then accrue mutations at the host neutral rate of evolution since they are fixed in the host genome (Katzourakis, 2013). Non-functional EVEs are expected to accumulate mutations at a far slower rate than their exogenous viral counterparts (Aiewsakun and Katzourakis, 2015). EVEs that have been functionally co-opted by the host cell would be expected to have an even slower mutation rate due to being conserved through positive selection. Demographic patterns such as host or viral population size could also affect the viral endogenization. Host species with small effective population sizes (i.e., many mammal species) may contain more neutral EVEs due to the amplified importance of genetic drift (Holmes, 2011).

In recent years, several studies have illustrated how exaptation of EVEs into a host's genome function in antiviral defense through production of functional proteins (Frank and Feschotte, 2017). For example, in the thirteen-lined ground squirrel (*Ictidomys tridecemlineatus*), an endogenous bornavirus effectively stops infection of exogenous (i.e., viruses which exhibit horizontal transmission from organism to organism) bornaviruses (Fujino et al., 2014). EVEs can have an influential and enduring influence on their host's evolution and subsistence in a given ecosystem. Therefore, analyzing EVEs across host genomes could help to unravel the complicated evolutionary relationships between viruses and their hosts.

Preliminary evidence suggests that ant genomes contain several clades of EVEs. François et al. (2016) found several putative ant EVE hits when surveying the Parvoviridae viral clade. Li et al. (2015) found a few ant glycoprotein EVEs in a study aimed at discovering arthropod RNA viruses. Dennis et al. (2018) found a single *Pseudomyrmex gracilis* cyclovirus ant EVE in a large survey of Circoviridae EVEs. However, these three studies used a virus-centered approach and only incidentally discovered ant EVEs. Their approach targets a small group of viruses among a wide array of host genomes in order to understand more about that clade of viruses. Conversely, a host-centered approach, in which one surveys a specific group of host genomes for all known viruses, permits the discovery of novel EVEs within those hosts.

Ant species exhibit extremely variable diets (herbivore, predator, and generalist), nesting habitat (arboreal vs. ground), colony structure, and complex and species-specific social behavior (Hölldobler and Wilson, 1990; Lach et al., 2010). Examination of ant EVEs may provide insight into this variability across their evolutionary history. Though all ants share the same RNAi immune response pathway, differences in EVEs across species may signify differential viral pathogen infection rates (Mongelli and Saleh, 2016). Therefore, analysis of EVEs could provide understanding of insect immunity evolution by acting as a reservoir of immune memory. In addition, a survey of EVEs throughout ant genomes could help elucidate the factors shaping the composition of viral communities presently infecting ants. EVEs scattered throughout the genomes of ants could represent a deep branch of the antiviral defense system (Whitfield et al., 2017).

Though there are currently nineteen published ant genomes, there have been no genome-wide studies examining their endogenous viruses. Therefore, the goal of this study is to survey and characterize EVE sequences throughout these nineteen ant genomes. Specifically we aim to address three questions: (1) Do ants exhibit abundant and diverse EVEs throughout their genomes? (2) How are the EVEs found in ant genomes related to exogenous viral clades? (3) Do any of these discovered EVEs exhibit potential for functionality?

MATERIALS AND METHODS

A comprehensive bioinformatic pipeline using BLAST was created to screen for EVEs throughout every published ant genome in the NCBI database[1]. There are currently nineteen published ant genomes (**Table 1**). These assembled genomes are of various sizes ranging from 212.83 megabases to 396.25 megabases. Before each ant genome was screened, scaffolds under 10,000 base pairs (bp) were pruned from the genome with the program CutAdapt (Martin, 2011) to ensure EVE hits were located on the actual genome and not scaffolds potentially created through assembler error or contamination.

The bioinformatic screen took a conservative approach, which consisted of first executing tblastn on RefSeq viral proteins (and all proteins from Shi et al., 2016) as the query against the specific ant genome as the database. The *e*-value for this tblastn was set at 1e−20 (Blast., 2013). Viral proteins from Shi et al. (2016) were included in the query because this study vastly increased the known viral diversity in insects and was not included in the RefSeq database at the time of this analysis. At the time of the screen, there were 1,149,421 viral proteins included and nineteen ant genomes analyzed. The nucleotide hits from this tblastn were merged with neighboring hits within 10 base pairs into a single sequence. These merged hits were then used as the query for a blastx run against the non-redundant protein database. The *e*-value for this blastx run was set at 0.001. The purpose of this blastx run was to assess if the original hit from the tblastn run was of viral origin. If the best hit was not most similar to a virus, then it was discarded. The final list of putative amino acid EVE hits were then manually pruned to ensure the best hit was also not most similar to a hypothetical protein, but to a viral structural or non-structural protein. Several EVEs were manually concatenated if they were close to each other on the same scaffold and when aligned did not overlap, but instead came from one larger protein fragment.

Once these putative EVEs were identified, the phylogenetic relationships of the EVE hits were inferred from an amino acid alignment of EVE protein hits and their closely related exogenous virus protein sequences, which were determined by most similar BLAST match. EVEs were grouped together based on both the viral protein class (glycoprotein, RNA-dependent

[1] https://www.ncbi.nlm.nih.gov/; accessed 1/1/2018

RNA polymerase, nucleoprotein, etc.) and viral clade to which it was most similar on the blastx run. These EVEs and their closely related exogenous viral protein sequences were aligned with MAFFT v7.309 (Katoh et al., 2002). For each alignment, the MAFFT program used the algorithm E-INS-I, scoring matrix of BLOSUM62, and a gap open penalty of 1.53. Subsequently, maximum likelihood phylogenies were inferred using the amino acid alignments with RAxMLv8.1.16 (Stamatakis, 2014). The best fit protein substitution model for each alignment was determined using SMS (Lefort et al., 2017). Support for the maximum likelihood (ML) phylogenies was evaluated with 1000 bootstrap replicates.

We assessed if ant genome quality was correlated with the number of EVEs present in the genome. For these analyses, we assessed both the genome directly downloaded from NCBI (pre-clipped) as well as the clipped genome used in the pipeline. By using the Pearson's product-moment correlation, we compared number of EVEs present in each genome with genome length, number of scaffolds, scaffold N50, number of contigs, and contig N50. BBMap was used to compile the statistical metrics for the clipped genomes (Bushnell, 2014). To understand if size filtering impacted the EVEs we found, we performed a synteny analysis to assess the size and number of annotated host genes on the individual scaffolds in which EVEs were discovered. This analysis was manually performed by examining each scaffold in the NCBI Genome Data Viewer.

To further examine EVE-ant evolutionary relationships, we used BaTS Bayesian tip-association significance testing (Parker et al., 2008; Shi et al., 2018) to assess if the EVE hits from the Mono-Chu glycoprotein phylogeny tend to clump more strongly with a particular ant species than expected solely by chance. The Mono-Chu glycoprotein phylogeny was simplified to include only the 227 ant EVE hits since we were only testing ant host-EVE associations. This test considered host phylogenetic structure at the level of ant species and subfamily. BaTS then estimated an association index (AI) to identify the strength of the association between the Mono-Chu glycoprotein EVE phylogeny and ant host species/subfamily. This AI value was then compared to a null distribution generated using 1000 tree-trip randomizations to infer an Association Index Ratio (observed association index/null association index). A ratio closer to 0 suggests stronger host structure and closer to 1 suggests a weaker host structure. A *p*-Value for the AI was output from the BaTS test derived from the 1000 tree-tip randomizations.

To evaluate the degree of EVE-ant host co-divergence in each ant species, we implemented an event-based co-phylogenetic reconstruction using JANE version 4 (Conow et al., 2010). The simplified EVE hit Mono-Chu glycoprotein phylogeny used for the BaTS test was used for this analysis. In addition, for the host phylogeny we used the ant phylogeny from Nelsen et al. (2018), with the drop.tip function in *ape* to infer a phylogeny with solely the 19 ant species examined in this study. The cost event scheme, or non-co-divergence, for the JANE reconstruction was: co-divergence = 0, duplication = 1, host switch = 1, loss = 1, failure to diverge = 1. The "failure to diverge" parameter refers to occurrences when host speciation is not followed by virus speciation, and the virus remains on both newly speciated hosts. The population size and the number of generations were fixed to 100. The co-divergence significance was determined by contrasting the estimated costs to null distributions calculated from 100 randomizations of host tip mapping. To better visualize these co-divergence patterns, we visualized these associations between the EVEs in the simplified glycoprotein Mono-Chu phylogeny and the EVEs in the simplified Nelsen et al. (2018) ant phylogeny using the cophylo function in phytools to create a tanglegram (Revell, 2012).

The potential functionality of these endogenous viral fragments was assessed through the analysis of the stop codons and nonsense mutation within the EVE hit protein fragments to determine if they possess intact open reading frames (ORFs). Intact ORFs were then inferred to be functional if through manual comparison, the putative EVE protein sequence was within 75 amino acids in length of its most closely related exogenous viral protein.

RESULTS

Using a host-centered approach, we screened all 19 available and published ant genomes (on 1/1/2018) with a bioinformatics pipeline detailed in the Material and Methods section above. Once EVE hits were obtained, we assessed their genome-specific differences in abundance and variation. We recovered a total of 434 EVE hits across these 19 ant genomes (**Supplementary Tables S1**, **S2**). There is clear variation in the number of EVEs recovered depending on ant species (**Figure 1** and **Table 1**). **Table 2** displays the results of the Pearson's product-moment correlation, comparing the factors representing genome quality with EVE number per genome. Based on the Pearson's product-moment correlation, none of these factors were significantly correlated with number of EVEs discovered. **Supplementary Table S3** contains all genome information used for both the clipped and pre-clipped genomes. For the synteny analysis, we found that 78.57% of the scaffolds containing EVEs were longer than 30,000 base pairs. In addition, 70.74% of these scaffolds had at least one gene annotated from the host. The scaffold length and number of annotated host genes are found in **Supplementary Table S1**. There were no EVEs that represented an entire viral genome on a single scaffold – instead each EVE hit constituted a single protein or protein fragment.

Viral Phylogenies

A summary of information on the phylogenies presented in the results is shown in **Table 3**. All the phylogenies described in the subsequent results can be found in **Supplementary Figures S1–S23**.

ssRNA Viruses

For the names of the RNA viral clades we will be using the nomenclature defined by Shi et al. (2016). Due to the increase in diversity of RNA viruses discovered by Shi et al. (2016), a merger of different viral families and orders into larger super-clades was proposed. For example, the Bunya-Arena clade is comprised of all the viruses from Bunyaviridae, Tenuivirus, Arenaviridae, and

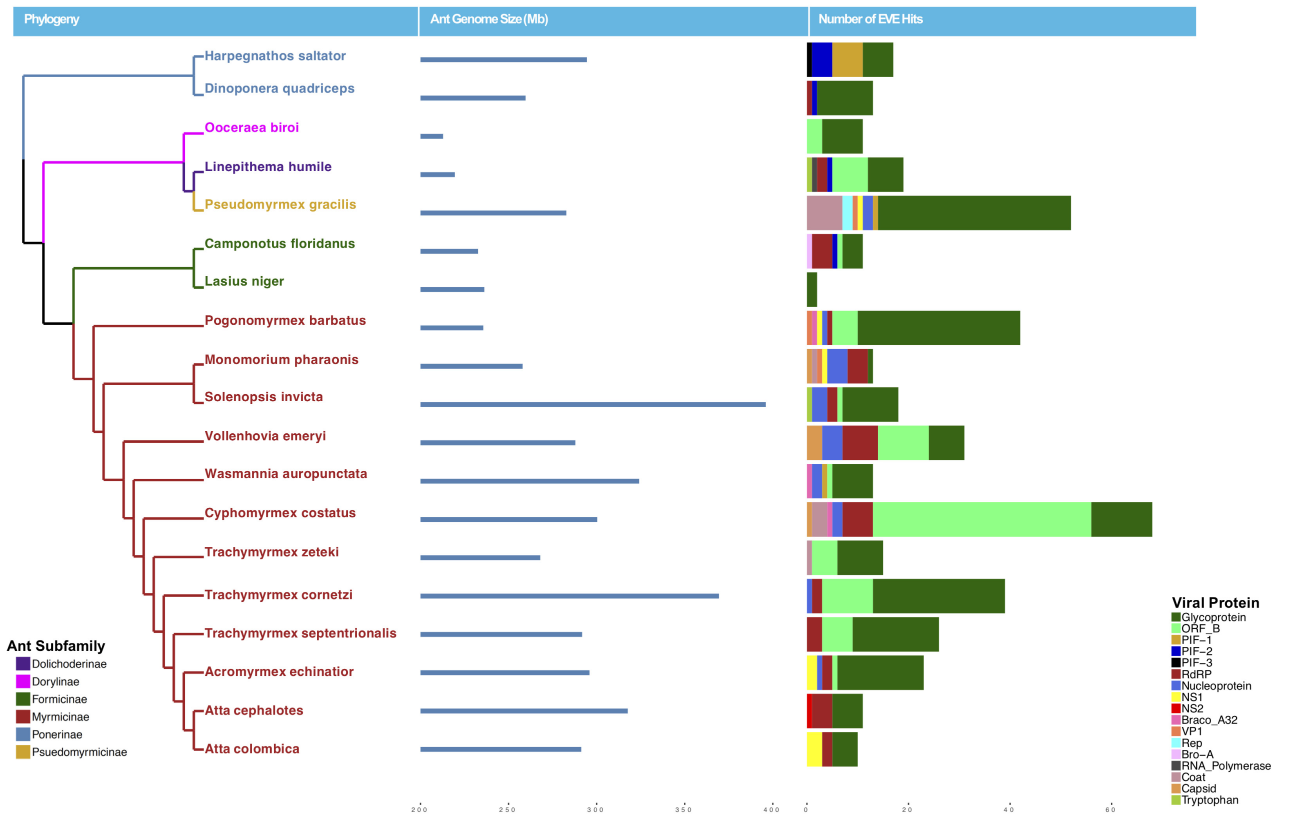


FIGURE 1 | Phylogeny of ants with sequenced genomes used in this study colored by subfamily (left), ant genome size in mb (middle), and number of EVE hits per genome (right). Ant EVE hit numbers are colored by viral protein to which they were found to be most similar in the blastx analysis.

TABLE 1 | Summary table of ant genomes which includes information on the species, subfamily, accession number from www.NCBI.nlm.nih.gov, total genome length in megabases, nesting habitat (arboreal/ground), diet (fungus, generalist, predatory, and herbivore), and number of EVE hits recovered.

Ant species	Ant subfamily	Accession number	Total length (Mb)	Nesting habitat	Diet	No. putative EVEs Recovered
Acromyrmex echinatior	*Myrmicinae*	GCF_000204515.1	295.945	Ground/arboreal	Fungus	23
Atta cephalotes	*Myrmicinae*	GCF_000143395.1	317.672	Ground	Fungus	11
Atta colombica	*Myrmicinae*	GCF_001594045.1	291.258	Ground	Fungus	10
Camponotus floridanus	*Formicinae*	GCF_000147175.1	232.685	Arboreal	Generalist	11
Cyphomyrmex costatus	*Myrmicinae*	GCF_001594065.1	300.317	Ground	Fungus	68
Dinoponera quadriceps	*Ponerinae*	GCF_001313825.1	259.666	Ground	Predatory	13
Harpegnathos saltator	*Ponerinae*	GCF_000147195.1	294.466	Ground	Predatory	17
Lasius niger	*Formicinae*	GCA_001045655.1	236.236	Ground	Generalist	2
Linepithema humile	*Dolichodirinae*	GCF_000217595.1	219.501	Ground/arboreal	Generalist	19
Monomorium pharaonis	*Myrmicinae*	GCF_000980195.1	257.977	Ground	Generalist	13
Ooceraea biroi	*Dorylinae*	GCF_000611835.1	212.826	Ground	Predatory	11
Pogonomyrmex barbatus	*Myrmicinae*	GCF_000187915.1	235.646	Ground	Herbivore	42
Pseudomyrmex gracilis	*Pseudomyrmicinae*	GCF_002006095.1	282.776	Arboreal	Generalist	52
Solenopsis invicta	*Myrmicinae*	GCF_000188075.2	396.025	Ground	Generalist	18
Trachymyrmex cornetzi	*Myrmicinae*	GCF_001594075.1	369.438	Ground	Fungus	39
Trachymyrmex septentrionalis	*Myrmicinae*	GCF_001594115.1	291.747	Ground	Fungus	26
Trachymyrmex zeteki	*Myrmicinae*	GCF_001594055.1	267.973	Ground	Fungus	15
Vollenhovia emeryi	*Myrmicinae*	GCF_000949405.1	287.901	Ground	Predatory	31
Wasmannia auropunctata	*Myrmicinae*	GCF_000956235.1	324.12	Ground/arboreal	Generalist	13

TABLE 2 | Genome quality assessment of factors from the clipped and pre-clipped genomes as correlated with the number of EVEs per genome.

Genome type	Factor	*r*	*p*-Value
Pre-clipped	Genome length	+0.211	0.386
	No. contigs	−0.105	0.669
	Contig N50	+0.415	0.078
	No. scaffolds	−0.160	0.515
	Scaffold N50	−0.177	0.467
Clipped	Genome length	+0.358	0.132
	No. contigs	+0.125	0.611
	Contig N50	−0.101	0.680
	No. scaffolds	−0.124	0.613
	Scaffold N50	+0.091	0.712

The different genome factors tested included total genome length (Mb), number of contigs, contig N50, number of scaffolds, and scaffold N50. The r and p-Values derive from the Pearson's product moment correlation.

Emaravirus. For more detailed information on which smaller viral groups comprise which super-clades, refer to Figure 2 of Shi et al. (2016).

Five different types of proteins found in ssRNA viruses were most similar to the EVE hits: glycoproteins, RNA-dependent RNA polymerases, nucleoproteins, capsid proteins, and coat proteins. All of these proteins have been previously discovered in insect virus genomes (Shi et al., 2016). Glycoproteins were the most commonly found in ant genomes, though only in the Mono-Chu clade. RNA-dependent RNA polymerases (RdRP) were discovered in every clade of ssRNA virus with EVE hits. A few EVE hits similar to nucleoproteins, capsid proteins, and coat proteins were also discovered within distinct ssRNA viral clades. Viral clade phylogeny results are presented in alphabetical order.

Bunya-Arena

A total of 17 ant EVEs are most closely related to the exogenous Bunya-Arena viral clade: 16 are most similar to Bunya-Arena nucleoprotein protein fragments and one is most similar to a Bunya-Arena RdRP protein fragment. In the nucleoprotein Bunya-Arena phylogeny, the 16 EVEs are distributed across five clades throughout the phylogeny, though several of these clades are not well-supported and therefore not conclusive (**Supplementary Figure S1**). Two *P. gracilis* EVEs are in a well-supported clade that is sister to a clade of bunyaviruses (bootstrap value = 100). Three *Vollenhovia emeryi* EVEs hits all cluster together in a weakly supported clade sister to the remaining EVE clades (bootstrap value = 38). Five ant EVEs cluster around Wuhan insect virus 16 in a weakly supported clade (bootstrap value = 16) and five other ant EVEs form a well-supported clade of their own (bootstrap value = 89). Also within this clade, *Monomorium pharaonis* EVE10 is weakly recovered as sister to a Topsovirus clade, which primarily consists of plant viruses (bootstrap value = 29). In the RdRP Bunya-Arena phylogeny, the EVE hit, *Cyphomyrmex costatus* EVE20 is sister to Wuhan insect virus 16 in a well-supported clade (bootstrap = 100; **Supplementary Figure S2**).

Hepe-Virga

In this RdRP Hepe-Virga phylogeny, the three Hepe-Virga-like ant EVEs are all from Myrmicinae genomes and form a well-supported clade with Hubei virga-like virus 1 and 2 (bootstrap value = 100; **Supplementary Figure S3**).

Mono-Chu

Mono-Chu viruses are the most common viral clade of EVEs within all ant genomes (total of 253 EVEs), primarily due to glycoprotein representing 89.7% of Mono-Chu EVE hits. The other EVE hits are most similar to Mono-Chu nucleoproteins and RdRP. These EVEs came from all 19 ant genomes observed in this study. The Mono-Chu Glycoprotein phylogeny was described in a comprehensive manner within these results due to its use in subsequent analyses. In this phylogeny, there are a total of 21 clades that include ant EVEs (**Figure 2A** and **Supplementary Figure S4**). To examine this phylogeny in greater detail refer to http://itol.embl.de/shared/pflynn.

Two clades within this Mono-Chu glycoprotein phylogeny cluster with exogenous viral lineages. Clade 1 is moderately supported and consists of Hubei chuvirus-like virus 1 clustering with two EVE hits (bootstrap value = 55). The other clade is Clade 21, which consists of Hubei diptera virus 11 and two EVE hits (bootstrap value = 94).

The rest of the 224 ant EVEs all fall within ant EVE-specific clades that are not closely related to any exogenous viruses shown in **Figure 2B** (bootstrap value = 64). Within this part of the phylogeny there are 19 distinct clades, with all but four (Clade 10, 16, 19, and 20) having bootstrap support of 60 or higher. Clade 2 consists of 14 ant EVEs and is moderately well-supported (bootstrap value = 63). Clade 3 is distinct with an extremely well-supported long branch and consists of 13 EVEs (bootstrap value = 100). Clade 4 is highly supported and consists entirely of ten closely related *P. gracilis* EVEs (bootstrap value = 100). Clade 5 is a distinct well-supported clade and consists of 14 EVEs (bootstrap value = 98). Clade 6 is highly supported and 35 EVEs form this clade (bootstrap value = 97). Ten EVEs in this clade cluster together and are all *Pogonomyrmex barbatus* EVEs. Seven EVEs from the *Trachymyrmex* species group together, five *Cyphomyrmex costatus* EVEs group together, five EVEs from *P. gracilis* cluster, and eight EVEs exhibit no host-specific pattern.

Clade 7 is a moderately well supported clade and consists of 23 ant EVEs (bootstrap value = 77). The distinct Clade 8 consists of six EVEs exclusively from the Ponerinae subfamily (bootstrap value = 93). Clade 9 consists only of *Ooceraea biroi* EVEs (bootstrap value = 78). *P. barbatus* EVE25 is a single EVE that forms a not well-supported Clade 10 (bootstrap value = 21). Clade 11 consists of three EVEs from *Linepithema humile* (bootstrap value = 78). The moderately well-supported Clade 12 comprises nine *P. barbatus* EVEs (bootstrap value = 72). Clade 13 is a well-supported clade which consists of 16 *P. gracilis* EVEs (bootstrap value = 91). Clade 14 is a well-supported clade consisted of 48 EVEs from fungus-growing ant genomes (bootstrap value = 81). Clade 15 is a distinct clade which consists of two EVEs from the subfamily Myrmicinae (bootstrap value = 90). Clade 16 consists of a single EVE: *Camponotus floridanus* EVE10 which is sister to the fungus-growing ant Clade 14 (bootstrap value = 37).

TABLE 3 | Summary table of viral phylogeny information.

Virus genome structure	Virus clade	Protein	No. of EVEs	No. of exogenous proteins used	Supplementary Figure
ssRNA	Bunya-Arena	Nucleoprotein	18	18	**Supplementary Figure S1**
		RdRP	3	12	**Supplementary Figure S2**
	Hepe-Virga	RdRP	3	15	**Supplementary Figure S3**
	Mono-Chu	Glycoprotein	227	17	**Supplementary Figure S4**
		Nucleoprotein	2	12	**Supplementary Figure S5**
		RdRP	24	30	**Supplementary Figure S6**
	Narna-Levi	RdRP	2	15	**Supplementary Figure S7**
	Partiti-Picobirna	Capsid	5	7	**Supplementary Figure S8**
		RdRP	4	16	**Supplementary Figure S9**
	Qinvirus	RdRP	3	10	**Supplementary Figure S10**
	Toti-Chryso	Coat	12	12	**Supplementary Figure S11**
		RdRP	1	15	**Supplementary Figure S12**
ssDNA	Circoviridae	rep-associated protein	2	10	**Supplementary Figure S13**
	Parvoviridae	VP1	3	10	**Supplementary Figure S14**
		Non-structural protein 1	8	18	**Supplementary Figure S15**
		Non-structural protein 2	1	12	**Supplementary Figure S16**
dsDNA	Baculoviridae	Bro-a	1	12	**Supplementary Figure S17**
		PIF-1	8	18	**Supplementary Figure S18**
		PIF-2	7	14	**Supplementary Figure S19**
		PIF-3	1	12	**Supplementary Figure S20**
	Poxviridae	Tryptophan	2	6	**Supplementary Figure S21**
		RNA-polymerase RP0147	1	7	**Supplementary Figure S22**
	Polydnaviridae	Pox A32	3	6	**Supplementary Figure S23**
ssRNA(RT)	Metaviridae	ORF B	93	NA	NA

This table includes the virus genome structure, the virus clade and protein to which the EVE hit was most similar, the number of EVEs found per protein phylogeny, the number of exogenous viral proteins used in the phylogenetic reconstruction, and the Supplementary Figure Information. There is no phylogeny for the EVE hits most similar to ssRNA(RT).

All six EVEs which form Clade 17 are from the *P. gracilis* genome (bootstrap value = 95). Clade 18 consists of two EVEs from *P. barbatus* (bootstrap value = 97). Clade 19 is a not well-supported clade of 12 EVEs in the Myrmicinae subfamily (bootstrap value = 18). Clade 20 consists of *D. quadriceps* EVE4, (bootstrap value = 37) which is sister to Clades 8–17.

In the Mono-Chu nucleoprotein phylogeny, two EVEs cluster together with Hubei chuvirus-like virus 1 in a well-supported clade (bootstrap value = 99; **Supplementary Figure S5**).

Eighteen ant EVEs fall into the Mono-Chu RdRP phylogeny (**Supplementary Figure S6**). Five of these EVEs are nested within a well-supported clade with Orinoco virus (bootstrap value = 93). Two EVEs belong to a well-supported clade with Berant Virus (bootstrap value = 100). *V. emeryi* EVE12 clusters with Beihai rhabdo-like virus 1, a virus infecting Crustaceans in a well-supported clade (bootstrap value = 100). *M. pharaonis* EVE4 belongs in a clade with two vertebrate bornaviruses. *V. emeryi* EVE1 is nested within a clade with Marburgvirus (bootstrap value = 10). Eight EVEs all derived from fungus-growing ant genomes are clustered in a well-supported clade with Shuangao Fly Virus 2 (bootstrap value = 96).

Narna-Levi

The two Narna-Levi-like ant EVEs are found in the *L. humile* genome. In the RdRP Narna-Levi phylogeny, these two *L. humile* EVEs are nested within a well-supported clade of narna-like viruses from insect hosts (bootstrap value = 96; **Supplementary Figure S7**).

Partiti-Picobirna

A total of nine ant EVEs are most similar to Partiti-Picobirna viruses related to the RdRP and Capsid proteins. *M. pharaonis* and *V. emeryi* genomes produced hits which are found within both of these protein phylogenies (**Supplementary Figures S8**, **S9**).

In the Partiti-Picobirna capsid phylogeny, the five ant EVEs belong to three distinct lineages (**Supplementary Figure S8**). The first is a well-supported clade which includes three *V. emeryi* EVEs clustering together (bootstrap value = 100) and are sister to Beihai barnacle virus 12 (bootstrap value = 99). The second lineage consists of *M. pharaonis* EVE9 as sister to Wuhan cricket virus 2 (bootstrap value = 85). The third lineage places *C. costatus* EVE12 as sister to Hubei tetragnatha maxillosa virus 8, a virus from an arachnid host, though this position is not as well-supported (bootstrap value = 57).

In the Partiti-Picobirna RdRP phylogeny, the four ant EVEs again fall into three distinct lineages (**Supplementary Figure S9**). Two EVEs cluster (bootstrap value = 100) and form a clade with Hubei partiti-like virus 29. *D. quadriceps* EVE9 falls out into well-supported clade of Partiti-Picobirna viruses from hosts

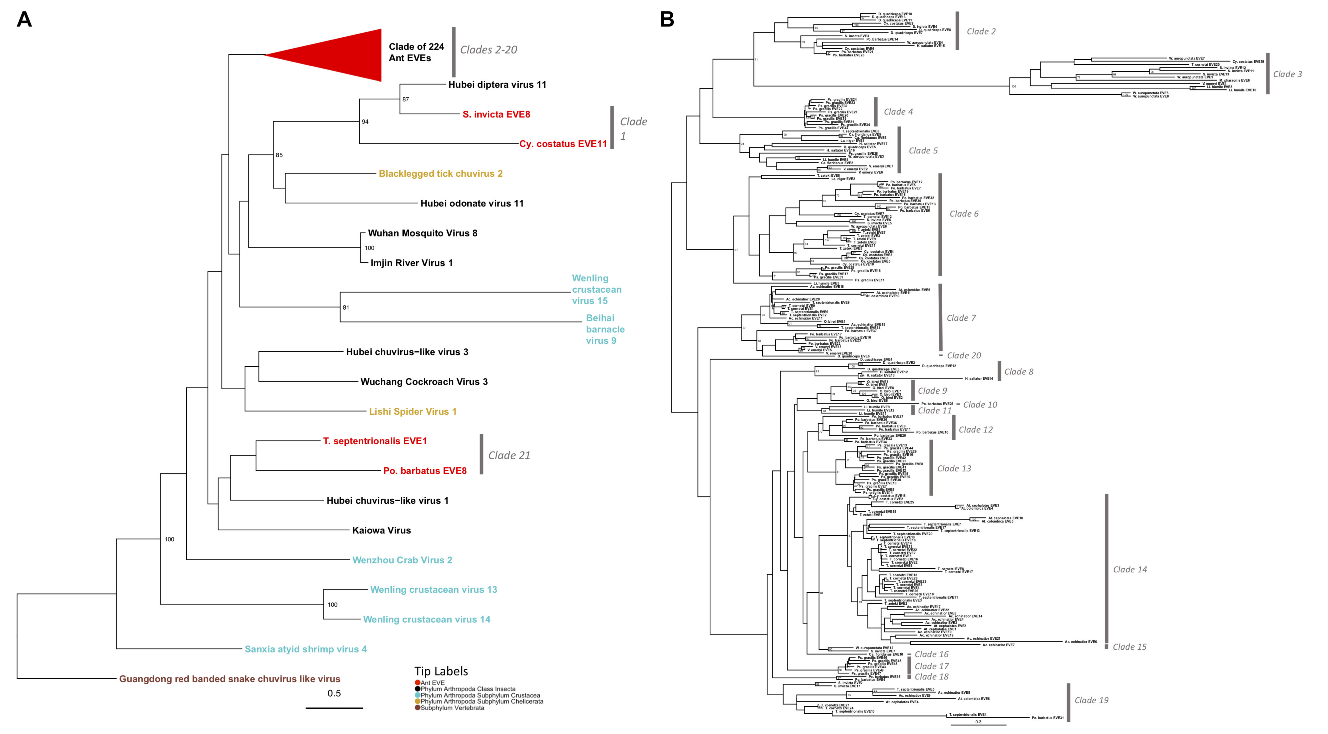


FIGURE 2 | (A) Glycoprotein Mono-Chu Phylogeny. The phylogeny was obtained from maximum likelihood analysis of the glycoprotein multiple amino acid alignment, including both ant EVE sequences and closely related exogenous viruses. The best fit protein substitution model was JTT + G + F. ML non-parametric bootstrap values >70 (1000 replicates) are indicated at each node. Scale bar indicates evolutionary distance in substitutions per site. The 224 ant EVEs which comprise their own clade have been collapsed and are represented by the red triangle. The tips are colored by associated host. The gray bars on the right are based on distinct clades of EVE hits. **(B)** Expanded phylogeny representing the clade of 224 ant EVEs which was collapsed in panel **(A)**. ML non-parametric bootstrap values >70 (1000 replicates) are indicated at each node. Scale bar indicates evolutionary distance in substitutions per site. The gray bars on the right are based on distinct clades of EVE hits.

of *Vespa velutina* (Asian hornet), Coleoptera (beetles), and Lophotrochozoa (snails) (bootstrap = 99). *Solenopsis invicta* EVE9 belongs to a lineage of Partiti-Picobirna viruses with hosts of both insect and chelicerate origin.

Qinvirus

The Qinvirus clade was first described in Shi et al. (2016) since the RdRP domains of the discovered viruses were so divergent from any previously known viral clade. *C. floridanus* EVE7 was discovered as an RdRP protein fragment most similar to the Qinvirus clade. In the reconstructed Qinvirus phylogeny, this EVE is sister to Wuhan insect virus 15 (bootstrap value = 78), nested within this larger clade (**Supplementary Figure S10**).

Toti-Chryso

Twelve of the thirteen Toti-Chryso-like ant EVEs are most similar to coat proteins. Seven of these hits are found in the *P. gracilis* genome. From the coat protein Toti-Chryso phylogeny, four of these *P. gracilis* EVEs cluster together as a distinct clade and as a sister clade to Shuangao toti-like virus and Australian anopheles totivirus (bootstrap value = 84; **Supplementary Figure S11**). Four EVEs belong in a weakly-supported Shaungao toti-like virus/Australian anopheles totivirus clade (bootstrap value = 42). Two EVEs cluster within a well-supported clade with *Leptopilina boulardi* toti-like virus (bootstrap value = 93). The last two *P. gracilis* EVEs are clustered within a well-supported clade of ant viruses such as *Camponotus yamaokai* virus and *Camponotus nipponicus* virus (bootstrap value = 99).

From the reconstructed RdRP Toti-Chryso phylogeny, *C. costatus* EVE25 is sister to the *L. boulardi* toti-like virus (bootstrap value = 85; **Supplementary Figure S12**).

ssRNA(RT) Viruses

Retroviruses (Metaviridae)

There were 93 ant EVEs similar to the ORF B (ORFs B) gene of *Trichoplusia ni* TED virus. This is an endogenous retrovirus found within the moth species, *T. ni*, and ORF B is a gene similar to the pol domain in retroviral genomes (Friesen and Nissen, 1990; Terzian et al., 2001). We could not reconstruct a phylogeny for these 93 EVEs because the ORF B gene has not been found in any other retroviruses to date.

ssDNA Viruses

Circoviridae

Replication-associated proteins (Rep) are responsible for genome replication within the viral Circoviridae clade (Dennis et al., 2018). Though weakly supported, in the Rep protein phylogeny, these two EVE hits are most closely related to the Dragonfly associated cyclovirus 2 (bootstrap value = 30; **Supplementary Figure S13**). *P. gracilis* EVE3 was previously discovered by Dennis et al. (2018) and conclusively, through targeted sequencing, shown to be an EVE in the *P. gracilis* genome.

Parvoviridae

Parvovirus genomes consist of a gene encoding a structural capsid protein (VP) and a non-structural protein, either Rep or NS (Bergoin and Tijssen, 2010; François et al., 2016). Many of the discovered viruses in the Parvoviridae clade have been found in vertebrate-hosts, however based on arthropod diversity, arthropod-host Parvoviridae viruses should vastly outnumber vertebrate viruses. In total, 12 ant EVEs are most similar to the Parvoviridae viral clade. The three EVE hits from the VP1 phylogeny all cluster together in a well-supported clade with Densovirus SC1065 (bootstrap value = 98; **Supplementary Figure S14**). The eight EVE hits from the NS1 phylogeny cluster into two different groups. Five of these EVEs cluster together into their own well-supported clade which is sister to a mosquito densovirus clade (bootstrap value = 98; **Supplementary Figure S15**). The other three EVEs cluster with the Lupine feces-associated densovirus in a moderately supported clade (bootstrap value = 65). Within the NS2 phylogeny, *Atta cephalotes* EVE5 clusters away from most of the insect densoviruses, but still clusters within a well-supported clade of several arthropod densoviruses (bootstrap = 98; **Supplementary Figure S16**).

dsDNA Viruses

Baculoviridae

Per os infectivity factor genes (PIF) and Baculovirus Repeated ORFs (Bro) are two common genes found within baculoviridae genomes which aid in host infection (Kang et al., 1999; Kikhno et al., 2002; Gauthier et al., 2015). A total of 17 ant EVEs are most similar to viruses from the Baculoviridae clade. The majority of these Baculoviridae EVE hits come from the *Harpegnathos saltator* genome (64.7% or 11 EVE hits), and all but one EVE is most similar to PIF fragments (PIF-1, PIF-2, and PIF-3). These EVEs which are similar to PIF fragments matched closely with *Apis mellifera* filamentous virus.

Camponotus floridanus EVE3 is most similar to a Bro-a fragment from *Mamestra brassicae* multiple nucleopolyhedrovirus. However, in the Bro-a phylogeny, this ant EVE did not fall anywhere within the nucleopolyhedrovirus lineage, which could suggest it is in its own viral genus within Baculoviridae (**Supplementary Figure S17**).

In the PIF-1 phylogeny, the eight EVE hits form a single ant-EVE specific distinct clade which is sister to *A. mellifera* filamentous virus (bootstrap value = 84; **Supplementary Figure S18**). Similarly, in the PIF-2 phylogeny the seven ant EVEs cluster together in a distinct clade with *A. mellifera* filamentous virus as the closest relative (bootstrap value = 83; **Supplementary Figure S19**). In addition, in the PIF-3 phylogeny the one EVE (*H. saltator* EVE1) clusters with *A. mellifera* filamentous virus (bootstrap value = 99; **Supplementary Figure S20**).

Poxviridae

There are a total of three ant EVEs which are most similar to the Poxviridae protein fragments Tryptophan repeat gene and RNA polymerase RPO147 (Theze et al., 2013; Mitsuhashi et al., 2014). The two EVEs within the tryptophan repeat gene phylogeny are found in separate clades within different lineages of entomopoxvirus (bootstrap value = 41; **Supplementary Figure S21**) and the *S. invicta* EVE1 clusters with the *Mythimna separata* entomopoxvirus (bootstrap value = 27). *L. humile* EVE1 falls within the RNA polymerase RPO147 clade, however it did not fall near any of the exogenous Entomopoxvirus viruses within this phylogeny (**Supplementary Figure S22**).

Polydnaviridae

Three ant EVEs are most similar to the *Cotesia congregata* bracovirus within the Polydnaviridae viral clade from the protein Pox A32 fragment (Espagne et al., 2004). In this phylogeny, all three EVEs fall into a clade with the *C. congregata* virus (bootstrap value = 100; **Supplementary Figure S23**).

Stop Codon Analysis

Of the 238 EVE hits of the 434 ant EVEs we recovered do not contain random stop codons (**Supplementary Table S4**). Sixteen of these EVEs without nonsense mutations are comparable in length to the viral proteins to which they are most similar (**Table 4**). Therefore, these hits are considered potentially functional or recently acquired.

EVE-Host Evolutionary Association Analyses

From the BaTS analysis, the association index ratio for the glycoprotein Mono-Chu EVE phylogeny in relation to the ant host phylogeny at the level of ant species was 0.263 with a significant p-Value of <0.001 and at the level of ant subfamily with a AI ratio of 0.090 and a significant p-Value of <0.001. From the JANE analysis of EVE-ant species co-divergence, we found that there were 26–30 co-divergence events, 65–68 host switching events, 131–132 duplication events 10–17 extinction events, and 278 total cost (combination of all non-co-divergence events) with a significant p-Value for the number of costs at <0.01. The tanglegram visualization illustrates the associations between ant phylogeny and the ant EVE phylogeny (**Figure 3**).

DISCUSSION

EVE Diversity

This study has greatly expanded our knowledge of EVEs found within ants. EVE hits from ant genomes are derived from a strikingly diverse set of viral lineages both from RNA and DNA viruses. Overall, our phylogenetic analysis found that ant EVEs tend to group into distinct, well-supported clusters from more than 12 viral lineages. There are major differences in abundance in EVEs across ant genomes; *Lasius niger* has the fewest with only two EVEs whereas *C. costatus* has the largest number with 68 EVEs. This could reflect biologically distinct rates of endogenization by viruses into certain ant genomes. However,

TABLE 4 | Potentially functional stop codons.

EVE	Similar virus	Similar protein	EVE length (AA)	Protein length (AA)
Cyphomyrmex costatus EVE18	Wuchang cockroach virus 3	Glycoprotein	586	659
Linepithema humile EVE6	Hubei narna-like virus 19	RdRP	750	737
Monomorium pharaonis EVE2	Densovirus SC1065	VP1	215	288
Pogonomyrmex barbatus EVE34	Wuhan mosquito virus 8	Glycoprotein	582	653
Pseudomyrmex gracilis EVE3	Cyclovirus PK5034	Rep-associated	236	277
Pseudomyrmex gracilis EVE4	Densovirus SC1065	VP1	281	288
Pseudomyrmex gracilis EVE41	Wuchang cockroach virus 3	Glycoprotein	641	659
Pseudomyrmex gracilis EVE45	Wuhan insect virus 16	Nucleoprotein	294	327
Pseudomyrmex gracilis EVE48	Wuhan mosquito virus 8	Glycoprotein	624	653
Solenopsis invicta EVE17	Wuhan mosquito virus 8	Glycoprotein	582	653
Trachymyrmex cornetzi EVE19	Wuchang cockroach virus 3	Glycoprotein	605	659
Trachymyrmex septentrionalis EVE19	Wuhan mosquito virus 8	Glycoprotein	612	653
Vollenhovia emeryi EVE10	Hubei partiti-like virus 11	Capsid	412	475
Vollenhovia emeryi EVE17	Wuhan insect virus 16	Nucleoprotein	327	327
Vollenhovia emeryi EVE19	Wuhan insect virus 16	Nucleoprotein	326	327
Wasmannia auropunctata EVE12	Wuhan mosquito virus 8	Glycoprotein	595	653

Summary table of EVEs which did not contain stop codons and were similar in size to their most analogous protein. This table contains information on the EVE, the most similar virus to that EVE, the most similar protein to that EVE, the length of the EVE protein hit in amino acids, and the length of the most similar protein in amino acids.

differences in genome sequencing and assembly quality may also contribute to this difference, as this might affect the number of EVEs one is able to detect in the genome. For example, when using long-read assembly for the *Aedes aegypti* genome, Whitfield et al. (2017) were able to discover a large and diverse number of EVEs. The different ant genomes vary considerably in their assembly statistics (**Supplementary Table S3**).

Statistically, the number of EVE hits per ant genome are not significantly correlated with any of the various factors relating to genome quality (genome length, scaffold statistics, contig statistics) (**Figure 1** and **Table 2**). In addition, based on the synteny analysis, close to 80% of the scaffolds these EVE hits were found on were over 30,000 bp long and around 70% contained annotated host genes. This suggests that the quality of the genomes was not biasing these analyses in a substantial manner. However, even though no genome quality factors were significantly correlated, several were close to a significant correlation (contig N50 and genome length; **Table 2**). In addition, when we took the *L. niger* genome out of the Pearson's product moment correlation analysis for these borderline significant tests, the correlation substantially decreased. This indicates the *L. niger* genome might be of lower quality than the other ant genomes and that genome quality does in fact constrain and potentially impact the abundance of EVE hits discovered.

In addition to the differences in EVEs per ant genome, there were marked differences in the number of EVEs per viral protein. It is plausible that certain viral types are better suited for attaining germline integration (Horie et al., 2010). For example, viruses which cause chronic infections in their host would most likely contain more EVEs than viruses solely causing acute infection (Holmes, 2011). Some viruses have been able to develop mechanisms to evade or inactivate the host's cellular DNA repair machinery, which would allow them to integrate more successfully into the host genome. Several dsDNA viruses (adenoviruses) have been found to inactivate DNA repair proteins to evade excision (Weitzman et al., 2004; Lilley et al., 2007). Therefore, specific clades of endogenous viruses found in our study would potentially reach fixation more often in ant genomes depending on their strategy to evade excision.

Additionally, RNA viruses have higher rates of mutation and replication in comparison to DNA viruses, which lends itself to more frequent host switching and larger host species range (Sanjuan et al., 2010; Pompei et al., 2012). Host range of the viral clades can impact the distribution of EVEs across the phylogeny. For example, certain genes in betaretroviruses allow them to have broader host ranges than gammaretroviruses (Henzy and Johnson, 2013). Based on our results, viruses from the Mono-Chu RNA virus clade may have certain genes or strategies which enabled them to colonize and endogenize within every ant species analyzed.

Understanding the EVE Phylogenies

Among many of the ssRNA viral protein phylogenies (Toti-Chryso, Narna-Levi, and Bunya-Arena) most ant EVEs cluster together and form monophyletic clades, suggesting they come from distinct and possibly ant-specific viral lineages. Interestingly many of these ant EVE-specific clades are most closely related to other insect or arthropod clades. Furthermore, almost all of the ssRNA viral lineages contained EVEs most similar to RdRP. This may be because RdRP is the only known conserved sequence domain across all RNA viruses and there are strong selection pressures acting on viral polymerases, leading to a high degree of conservation of such proteins (Hughes and Hughes, 2007; Shi et al., 2016). Qinvirus is a new ssRNA viral group discovered in 2016 and *C. floridanus* EVE7 is the first ant EVE from this viral lineage.

In addition to discovering many new ant EVEs, several of the EVEs previously discovered were reconfirmed in this study:

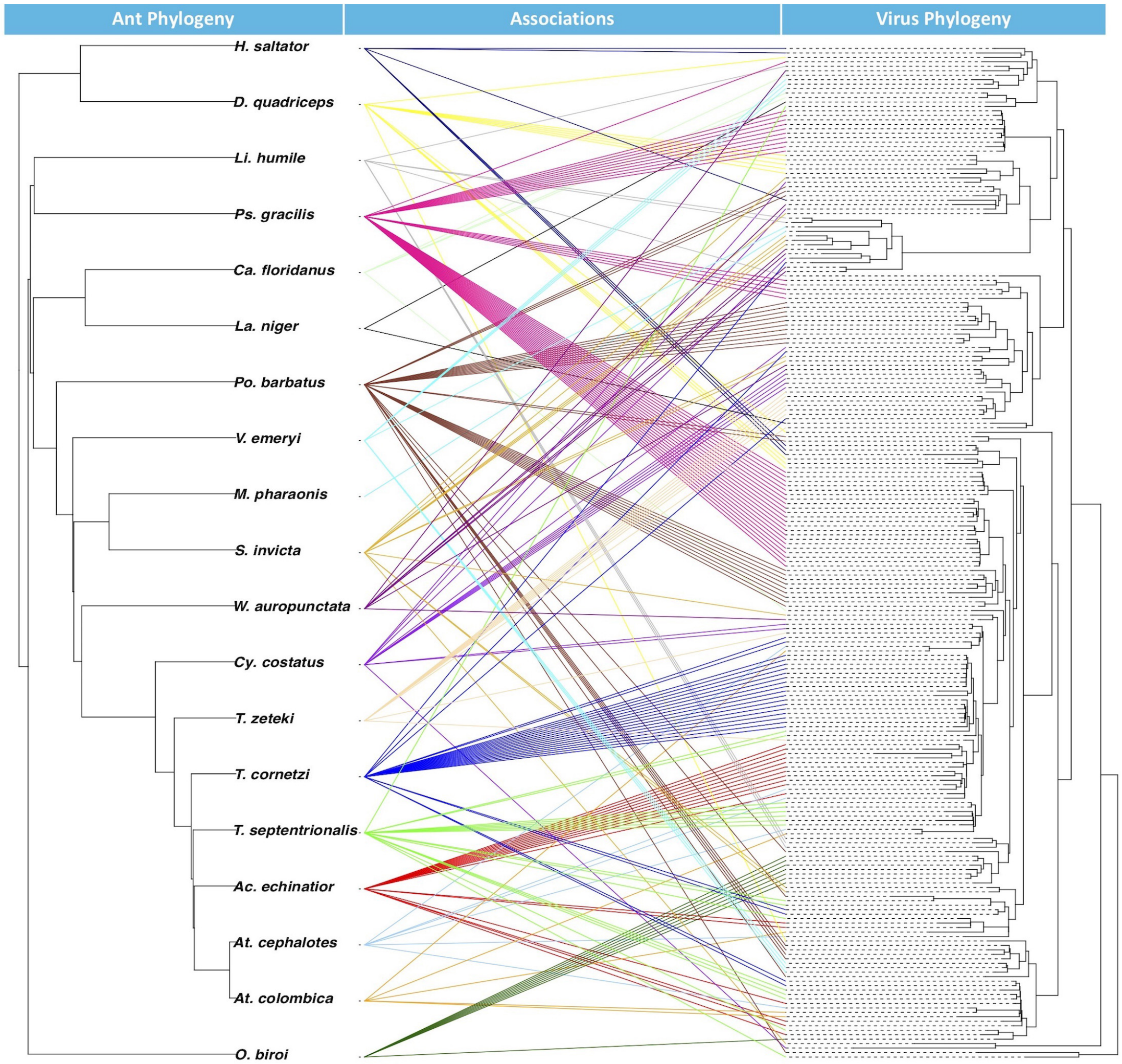


FIGURE 3 | Tanglegram with ant phylogeny on left and glycoprotein Mono-Chu viral phylogeny on right. The exogenous viruses included in the viral phylogeny were left out because their hosts are not included in the ant phylogeny. Colors of the association are randomly prescribed for each of the ant species.

P. gracilis EVE3 from the rep-association circoviridae phylogeny was previously found by Dennis et al. (2018). Three EVE hits (*Acromyrmex echinatior* EVE1,2 and *Monomorium pharaonis* EVE1) from the Parvoviridae NS1 phylogeny were previously found by François et al. (2016). This confirms that our methods were accurate in classifying EVEs.

Camponotus floridanus EVE3 in the Baculoviridae Bro-a protein phylogeny and *L. humile* EVE1 from the Baculoviridae RNA polymerase phylogeny did not fall out near any exogenous viruses. For this reason, these EVE hits may represent two new unknown viral clades. The three EVEs which were most similar to Polydnaviridae Pox A32 proteins are inferred to be most similar to the *C. congregata* virus, an endogenous wasp virus. This provides some evidence that these three lineages of ants could have been parasitized at some point over evolutionary history by a wasp species carrying this *C. Congregata* virus or alternatively that this virus has the potential to infect a diversity of Hymenoptera (ants, bees, and wasps). Similarly, over evolutionary history *H. saltator* ants could have been infected (and their genome endogenized) with an ancestor of the *A. mellifera* filamentous virus. This may be the reason that *H. saltator* genome contains several EVE protein fragments for PIF-1, PIF-2, and PIF-3; all proteins found within the *A. mellifera* filamentous virus.

One potential reason why there were so few retroviral EVEs discovered and so many RNA viruses discovered is an issue of sampling bias. There has been much more effort put into discovering RNA viruses within insects due to their impact on human populations (Dengue Virus, West Nile Virus, Zika Virus, and Yellow Fever Virus). Since, this pipeline can only identify the viruses available in the databases, a wide diversity and abundance of insect-specific RNA viruses were detected.

Retroviral ORF B EVE hits constituted around 21% of the total EVE hits for all ant genomes. Retroviral EVEs make up a large proportion of hits because endogenization into the host genome forms part of their replication cycle. However, it is still surprising that an even more diverse group of endogenous retroviruses were not found. One of the reasons we suspect that solely ORF B analogs were discovered for retroviruses is because there are only thirteen described insect-specific retroviruses. One of these thirteen described retroviruses is *T. ni* TED virus, which contains this ORF B gene (Terzian et al., 2001). These ORF B ant EVE hits are most similar to the moth endogenous retrovirus *T. ni* TED virus, which functions as a retrotransposon in the moth genome. This suggests that these EVEs could be classified as retrotransposons and should be further analyzed to better understand their evolutionary role in the ant genomes.

Many of the ant EVEs discovered in this study tend to cluster together in clades by ant species. Multiple EVEs from the same ant species form distinct clades seven times in the Mono-Chu glycoprotein phylogeny, twice in Toti-Chryso coat phylogeny, twice in the Bunya-Arena nucleoprotein phylogeny, once in the Partiti-Picobirna capsid phylogeny, and once in each of the Baculoviridae PIF-1 and PIF-2 phylogenies. This suggests that EVEs have often been generated either by multiple viral integration events or by one integration event and multiple duplication events. In addition, it implies that ant species are either more prone to persistent infection by viruses in certain lineages or that this viral sequence was repeatedly conserved in these specific ant genome over evolutionary time.

Mono-Chu Glycoprotein Analyses

We chose to focus much of the analyses on the Mono-Chu glycoprotein phylogeny as it was the only group where there were EVEs in all 19 ant genomes, therefore aiding in a more comparative analysis of the EVEs. One potential reason that glycoproteins constitute over half of the EVEs identified may be because viral glycoproteins are extremely important in viral infection and immunity. Glycoproteins often play a critical function in viral infection by identifying and binding to receptor sites on the host's membrane (Banerjee and Mukhopadhyay, 2016).

The 224 EVEs, which make up Clades 2–20 in the Mono-Chu glycoprotein phylogeny, all fall within a distinct ant-specific lineage not closely related to any exogenous viruses (bootstrap value = 64; **Figure 2B**). The considerable divergence between these ant EVEs and any other exogenous virus suggests that this lineage most likely represents a previously undiscovered clade of viruses. Since so few viruses currently infecting ants (and insects) have been identified, the additional resolution of this phylogeny with newly discovered viruses will help us understand if this Mono-Chu viral lineage consists of solely ant- or insect-specific viruses.

Since this Mono-Chu glycoprotein phylogeny contains so many ant EVEs, several interesting inferences can be made. Clade 14 consists solely of 48 fungus-growing ant EVEs. This clustering could be potentially due to infection with a virus specific to fungus-growing ants or alternatively, a prior infection in the ancestor of all fungus-growing ants. Clade 3 exhibits extremely long branch lengths compared to the other clades within this phylogeny. This implies that the infecting viruses endogenized fragments into the ant species in this clade longer ago in evolutionary time, giving this clade time to diverge from the rest of the viral fragments in this phylogeny, although there could also be faster rates of molecular evolution in this virus. As in many of the other inferred phylogenies, EVEs from certain ant species tend to clump together into distinct clades. This happens frequently in the Mono-Chu glycoprotein phylogeny. For example, *P. gracilis* EVEs cluster together into three distinct clades (Clades 4, 13, and 17). This suggests that *P. gracilis* might be more prone to viral infection by viruses from the Mono-Chu lineage than other ant species.

Examination of EVE-ant association reveals that though host switching is most likely common among these viruses, EVEs can mirror their ant hosts over evolutionary time within the Mono-Chu lineage. This is supported by the AI ratios, which found that the EVE phylogeny exhibited significant clustering ($p < 0.001$) by host taxonomy for both ant species and ant subfamily. The ant subfamily (0.090) exhibited much stronger structuring than for the ant species (0.263), however both of these values are much closer to 0 than to 1, suggesting that the ant species are strongly structuring the EVE phylogenetic relationships. The co-phylogenetic analysis was performed at the ant species level, and found significantly more EVE-ant host co-divergence than solely by chance ($p < 0.01$). Overall this implies a long-term association of Mono-Chu viruses and their ant hosts. However, host-switching of these EVEs seems to also be very frequent throughout their evolution, and even more common than co-divergence with ant species in the Mono-Chu viral glycoprotein lineage. Therefore, even though there looks to be a long-term association between viruses from the Mono-Chu lineage in ant genomes, cross-species host switching occurred commonly as the viruses/ants co-diverged. One can visually assess this association from the tanglegram, which illustrates topological incongruence that suggests host-switching, though the EVEs tend to clump by ant species implying potential for co-divergence (**Figure 3**).

Functionality

Around half of the EVEs we found across the ant genomes included nonsense mutations from premature stop codons. These stop codons tend to accumulate over evolutionary time in parts of the genome which are not functional. However, 238 EVE hits (54.8% of all EVEs discovered) did not contain stop codons (**Supplementary Table S4**) and are considered intact ORFs. In addition, sixteen of these hits were roughly the same size as the viral proteins they were most similar to, which implies that these sequences are still potentially functional within the ant genome (**Table 4**). Finding these potentially functional ORFs could either mean that there was a recent origin of the EVE or that the fragment was conserved over evolutionary time. If it is the latter, then these EVEs could potentially serve a current function in the genome, such as an antiviral defense mechanism co-opted by the ant. The genomes of certain ant species might also be predisposed to accumulating more functional or non-functional EVEs depending on population demographics and exposure to specific viruses over time.

CONCLUSION

Newly discovered EVEs were found within all ant genomes and are similar to a large diversity of viral lineages. Many of these viral lineages do not contain currently known exogenous viruses from ant hosts, although several are closely related to other insect and arthropod exogenous viruses. Certain ant genomes tend to contain more abundant EVEs within them. Many closely related EVEs tend to cluster by species, which suggests multiple integration or duplication events within ant species. In addition, through analysis of EVEs similar to viral glycoproteins, host switching appears to be common among EVEs found in ants, though many EVEs have long-term associations with ant species and ant subfamilies. Furthermore, the potential for functionality of several of these EVEs supports the idea that EVEs could be playing an important role in ant genomes.

AUTHOR CONTRIBUTIONS

PF and CM conceived, designed, and executed the study and revised the manuscript. PF analyzed the data and wrote the manuscript.

FUNDING

PF was supported by a National Science Foundation Graduate Research Fellowship. CM was supported by a National Science Foundation grant (NSF DEB 1900357).

ACKNOWLEDGMENTS

We thank Felix Grewe for computational biology assistance and the Field Museum of Natural History for access to their servers. We also thank Benjamin Rubin for his help in determining the bioinformatic pipeline for this project. We also thank Maureen Coleman, Michaela Gack, and Cathy Pfister for their helpful discussions. We also thank the National Science Foundation for support (NSF DEB 1900357).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01139/full#supplementary-material

REFERENCES

Aiewsakun, P., and Katzourakis, A. (2015). Endogenous viruses: connecting recent and ancient viral evolution. *Virology* 479–480, 26–37. doi: 10.1016/j.virol.2015.02.011

Banerjee, N., and Mukhopadhyay, S. (2016). Viral glycoproteins: biological role and application in diagnosis. *Virusdisease* 27, 1–11. doi: 10.1007/s13337-015-0293-5

Bergoin, M., and Tijssen, P. (2010). "Densoviruses: a highly diverse group of arthropod parvoviruses," in *Insect Virology*, eds S. Asgari and K. N. Johnson (Poole: Caister Academic Press).

Blast. (2013). *BLAST Basic Local Alignment Search Tool, Blast Program Selection Guide*. Bethesda MD: U.S. National Library of Medicine.

Bushnell, B. (2014). *BBMap: a Fast, Accurate, Splice-Aware Aligner*. Berkeley, CA: Lawrence Berkeley National Lab. (LBNL).

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol. Biol.* 5:16. doi: 10.1186/1748-7188-5-16

Dennis, T. P. W., Flynn, P. J., Souza, M., De Singer, J. B., Moreau, C. S., Wilson, S. J., et al. (2018). Insights into circovirus host range from the genomic fossil record. *J. Virol.* 92, 1–9. doi: 10.1128/JVI.00145-18

Espagne, E., Dupuy, C., Huguet, E., Cattolico, L., Provost, B., Martins, N., et al. (2004). Genome sequence of a polydnavirus: insights into symbiotic virus evolution. *Science* 306, 286–290. doi: 10.1126/science.1103066

Feschotte, C., and Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nat. Rev. Genet.* 13, 283–296. doi: 10.1038/nrg3199

François, S., Filloux, D., Roumagnac, P., Bigot, D., Gayral, P., Martin, D. P., et al. (2016). Discovery of parvovirus-related sequences in an unexpected broad range of animals. *Sci. Rep.* 6:30880. doi: 10.1038/srep30880

Frank, J. A., and Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Curr. Opin. Virol.* 25, 81–89. doi: 10.1016/j.coviro.2017.07.021

Friesen, P. D., and Nissen, M. S. (1990). Gene organization and transcription of TED, a lepidopteran retrotransposon integrated within the baculovirus genome. *Mol. Cell. Biol.* 10, 3067–3077. doi: 10.1128/mcb.10.6.3067

Fujino, K., Horie, M., Honda, T., Merriman, D. K., and Tomonaga, K. (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13175–13180. doi: 10.1073/pnas.1407046111

Gauthier, L., Cornman, S., Hartmann, U., Cousserans, F., Evans, J. D., De Miranda, J. R., et al. (2015). The apis mellifera filamentous virus genome. *Viruses* 7, 3798–3815. doi: 10.3390/v7072798

Grasis, J. A. (2017). The intra-dependence of viruses and the holobiont. *Front. Immunol.* 8:1501. doi: 10.3389/fimmu.2017.01501

Henzy, J. E., and Johnson, W. E. (2013). Pushing the endogenous envelope. *Philos. Trans. R. Soc. B Biol. Sci.* 368:20120506. doi: 10.1098/rstb.2012.0506

Hölldobler, B., and Wilson, E. O. (1990). *The ants*. Cambridge, MA: Belknap Press of Harvard University Press.

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe* 10, 368–377. doi: 10.1016/j.chom.2011.09.002

Horie, M., Honda, T., Suzuki, Y., Kobayashi, Y., Daito, T., Oshida, T., et al. (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463, 84–87. doi: 10.1038/nature08695

Hughes, A. L., and Hughes, M. A. K. (2007). More effective purifying selection on RNA viruses than in DNA viruses. *Gene* 404, 117–125. doi: 10.1016/j.gene.2007.09.013

Kang, W., Suzuki, M., Zemskov, E., Okano, K., and Maeda, S. (1999). Characterization of baculovirus repeated open reading frames (bro) in *Bombyx mori* nucleopolyhedrovirus. *J. Virol.* 73, 10339–10345. .

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Katzourakis, A. (2013). Paleovirology: inferring viral evolution from host genome sequence data. *Philos. Trans. R. Soc. B Biol. Sci.* 368:20120493. doi: 10.1098/rstb.2012.0493

Katzourakis, A., and Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genet.* 6:e1001191. doi: 10.1371/journal.pgen.1001191

Kikhno, I., Gutiérrez, S., Croizier, L., Crozier, G., and López Ferber, M. (2002). Characterization of pif, a gene required for the per os infectivity of Spodoptera littoralis nucleopolyhedrovirus. *J. Gen. Virol.* 83, 3013–3022. doi: 10.1099/0022-1317-83-12-3013

Lach, L., Parr, C. L., and Abbott, K. L. (2010). *Ant Ecology*. Oxford: Oxford University Press.

Lefort, V., Longueville, J. E., and Gascuel, O. (2017). SMS: smart model selection in PhyML. *Mol. Biol. Evol.* 34, 2422–2424. doi: 10.1093/molbev/msx149

Li, C. X., Shi, M., Tian, J. H., Lin, X. D., Kang, Y. J., Chen, L. J., et al. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* 4:e05378. doi: 10.7554/elife.05378

Lilley, C. E., Schwartz, R. A., and Weitzman, M. D. (2007). Using or abusing: viruses and the cellular DNA damage response. *Trends Microbiol.* 15, 119–126. doi: 10.1016/j.tim.2007.01.003

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mitsuhashi, W., Miyamoto, K., and Wada, S. (2014). The complete genome sequence of the *Alphaentomopoxvirus Anomala cuprea entomopoxvirus*, including its terminal hairpin loop sequences, suggests a potentially unique mode of apoptosis inhibition and mode of DNA replication. *Virology* 452–453, 95–116. doi: 10.1016/j.virol.2013.12.036

Mongelli, V., and Saleh, M.-C. (2016). Bugs are not to be silenced: small RNA pathways and antiviral responses in insects. *Annu. Rev. Virol.* 3, 573–589. doi: 10.1146/annurev-virology-110615-042447

Nelsen, M. P., Ree, R. H., and Moreau, C. S. (2018). Ant–plant interactions evolved through increasing interdependence. *Proc. Natl. Acad. Sci. U.S.A.* 115, 12253–12258. doi: 10.1073/pnas.1719794115

Parker, J., Rambaut, A., and Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* 8, 239–246. doi: 10.1016/j.meegid.2007.08.001

Pompei, S., Loreto, V., and Tria, F. (2012). Phylogenetic properties of RNA viruses. *PLoS One* 7:e44849. doi: 10.1371/journal.pone.0044849

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *J. Virol.* 84, 9733–9748. doi: 10.1128/JVI.00694-10

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

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Terzian, C., Pelisson, A., and Bucheton, A. (2001). Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol. Biol.* 1:3. doi: 10.1186/1471-2148-1-3

Theze, J., Takatsuka, J., Li, Z., Gallais, J., Doucet, D., Arif, B., et al. (2013). New insights into the evolution of Entomopoxvirinae from the complete genome sequences of four Entomopoxviruses infecting *Adoxophyes honmai*, *Choristoneura biennis*, *Choristoneura rosaceana*, and *Mythimna separata*. *J. Virol.* 87, 7992–8003. doi: 10.1128/jvi.00453-13

Weitzman, M. D., Carson, C. T., Schwartz, R. A., and Lilley, C. E. (2004). Interactions of viruses with the cellular DNA repair machinery. *DNA Repair* 3, 1165–1173. doi: 10.1016/j.dnarep.2004.03.018

Whitfield, Z. J., Dolan, P. T., Kunitomi, M., Tassetto, M., Seetin, M. G., Oh, S., et al. (2017). The diversity, structure, and function of heritable adaptive immunity sequences in the *Aedes aegypti* genome. *Curr. Biol.* 27, 3511.e–3519.e. doi: 10.1016/j.cub.2017.09.067

Conflict of Interest Statement: 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.

frontiers
in Microbiology

ORIGINAL RESEARCH
published: 25 June 2019
doi: 10.3389/fmicb.2019.01430

Adenovirus Isolated From a Cat Is Related to Human Adenovirus 1

Joseph Ongrádi[1,2]*, Louise G. Chatlynne[3]†, Katalin Réka Tarcsai[1], Balázs Stercz[1], Béla Lakatos[4], Patricia Pring-Åkerblom[5]‡, Donald Gooss Sr.[6]‡, Károly Nagy[1,2] and Dharam V. Ablashi[3]†

[1] Department of Medical Microbiology, Semmelweis University, Budapest, Hungary, [2] National Institute of Dermato-Venereology, Budapest, Hungary, [3] Advanced Biotechnologies Inc., Columbia, MD, United States, [4] Lakat-Vet BT, Budapest, Hungary, [5] Department of Virology, Hannover Medical School, Hanover, Germany, [6] Selbyville Animal Hospital, Selbyville, DE, United States

An adenovirus (AdV) has been isolated from the rectal swab of a domestic cat (*Felis catus*) and named feline adenovirus (FeAdV) isolate. It replicates and causes cytopathological effects in many human, feline, other mammalian cell lines that have both Coxsackie-adenovirus-receptor and integrins. Its antigens cross-react with anti-human adenovirus antibodies in immunofluorescence and immunocytochemistry assays. Electron microscopy revealed typical extracellular icosahedral particles and pseudo arrays inside cells. Sequence analysis of hexon and fiber genes indicates that this virus might belong to human adenovirus (HAdV) C species and might be a variant of type 1. In the fiber protein, three altered amino acids occur in the shaft; four altered residues are found in the knob region as compared to a European HAdV might be type 1 isolate (strain 1038, D11). One alteration affects amino acid 442 forming an RGS motif in an alanine rich region that might be an alternative way to bind integrins with subsequent internalization. Substitutions in the hexon sequence are silent. As compared to published HAdV sequences, the fiber is related to the original American prototype and recently described Taiwanese HAdV 1 isolates, but the hexon sequences are related to adenovirus isolates from France, Germany, Japan, and Taiwan. Serology carried out on FeAdV infected M426 cells indicates a prevalence of IgG in 80% of domestic cats in Delaware, United States. FeAdV isolate seems to be a recently recognized virus with possible pathogenic effects and, simultaneous human and feline infections are possible. Further molecular and biological characterization of this feline adenovirus isolate, as well as studies on both human and feline epidemiology and pathomechanisms, especially in endangered big cats, are warranted. FeAdV might have further practical advantages. Namely, it could be utilized in both human and feline AIDS research, developed into diagnostic tools, and gene therapy vectors in the near future.

Keywords: feline adenovirus isolate, HAdV-1 related, hexon sequence, fiber sequence, cell host range, interspecies transmission, emerging pathogen

OPEN ACCESS

Edited by:
Gustavo Caetano-Anollés,
University of Illinois
at Urbana–Champaign, United States

Reviewed by:
Donald Seto,
George Mason University,
United States
Jose Luiz Proenca-Modena,
University of Campinas, Brazil

***Correspondence:**
Joseph Ongrádi
ongradi.jozsef@
med.semmelweis-univ.hu;
ongjos@hotmail.com

†Present address:
Louise G. Chatlynne
and Dharam V. Ablashi,
HHV-6 Foundation, Santa Barbara,
CA, United States

‡Retired

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

Received: 06 December 2018
Accepted: 06 June 2019
Published: 25 June 2019

Citation:
Ongrádi J, Chatlynne LG, Tarcsai KR, Stercz B, Lakatos B, Pring-Åkerblom P, Gooss D Sr, Nagy K and Ablashi DV (2019) Adenovirus Isolated From a Cat Is Related to Human Adenovirus 1. Front. Microbiol. 10:1430. doi: 10.3389/fmicb.2019.01430

INTRODUCTION

Adenoviruses (AdV) are important, widespread [almost 100% of humans have circulating antibodies against the common human adenovirus (HAdV) types/see refs in Stercz et al., 2013], and occasionally fatal pathogens in humans, as well as wild and domestic animals. There are seven species (A through G) of Mastadenoviruses comprising a steadily increasing number of HAdV types. The HAdV Working Group has recognized 85 types recently[1] [The HAdV Working Group (2019)]. Among these, species C is the most important since it can establish latent infection as episomes in immune cells and has a high fatality rate in immunocompromised patients. Reactivation of latent AdVs after bone marrow (BMT) or hematopoietic stem cell (HSCT) transplantation might elicit lethal complications or life-long sequelae (Lion, 2014). Early genes of AdVs have been shown to transactivate human immunodeficiency virus (HIV) (Kliewer et al., 1989), and consequently, promote AIDS progression. Before introducing the highly active antiretroviral treatment (HAART), approximately 20% of patients died of untreatable gastroenteritis (Hierholzer, 1992). Species C types can be excreted in massive amounts in the stool for months after initial infection (Matthes-Martin et al., 2013). Recombinant AdVs (rAdVs) and their immunomodulatory effects have been widely studied as gene transfer vehicles. Modified or recombinant AdV obtained from animal sources could be less immunogenic and provide long-term expression of therapeutic genes (Stercz et al., 2013). Clinical importance and anthropozoonosis of adenovirus infection in major animal groups have not been elucidated. Despite apparent medical and veterinary importance, the pathomechanism has received little attention. One of the obstacles is that no ideal animal model exists for the simultaneous establishment of the underlying diseases and adenovirus latency with subsequent reactivation. Most AdVs studied have a very narrow *in vivo* and *in vitro* host range, although crossing animal host or tissue culture species barrier has been described (Lion, 2014).

The cat (*Felis catus*) as an experimental model has become important since the discovery of the feline immunodeficiency virus (FIV) and feline AIDS in 1986 (Pedersen et al., 1987) as the only natural small animal model for human AIDS (Elder et al., 2010). Beside anti-retroviral chemotherapy and vaccine trials, cats can be used to study the pathomechanism of interaction between retrovirus and several heterologous microbes through the course of feline AIDS (Ongrádi et al., 2013). AdV studies in cats or *Felidae* are hindered by the lack of data on natural infection and epidemiology, as well as research and diagnostic tools (Gonin et al., 1995). Early studies by electron microscopy suggested the presence of an AdV like virus particles in a black panther suffering from inclusion body hepatitis (Gupta, 1978) and a domestic cat suffering from disseminated adenovirus infection, feline leukemia virus (FeLV) infection and hepatic failure (Kennedy, 1993), but neither virus has been isolated nor has aspects of infection been studied. In the feline AIDS model, first, we studied field cats for possible natural infection. In the middle of 1990's we screened 470 European (from Hungary, Italy, Netherlands, and Scotland) cats for the presence of anti-AdV antibodies using HAdV-1 hexon antigen in a home-made ELISA, and found 9.8–20.3% seropositivity (Lakatos et al., 1996; Lakatos et al., 2000), in 162 American (CA) cats 26% positivity (Lakatos et al., 2000). Immunization by using purified HAdV-1 hexon protein induced a high titer of antibodies (>1:3200) in specific pathogen-free cats (Lakatos et al., 2000). In the second step, pharyngeal and rectal specimens from several seropositive domestic cats were screened for AdV DNA by nested PCR assay using consensus primers (Lakatos et al., 1997, 1999). The last step was to attempt isolation of a replication competent virus and its basic characterization. We managed to isolate an AdV from the PCR positive rectal specimen (Ongrádi, 1999). Virions were visualized by electron microscopy. The genes of hexon and fiber proteins, which determine AdV tropism and type specificity, were sequenced using species-specific and type-specific PCR. Gradually we have carried out biological studies. The range of permissive cells of human and animal origin was defined *in vitro*. Intracellular AdV antigens were shown by immunofluorescence (IFA) and immunocytochemistry (ICH) assays. Sequencing and biological probes suggest that the isolate is related to HAdV type 1. It is described as feline adenovirus (FeAdV) isolate. This isolate can also be numbered as FeAdV-1, as more isolates can be expected in the future.

MATERIALS AND METHODS

Data Availability

Publicly available datasets were analyzed in this study. This data can be found here: http://www.ncbi.nlm.nih.gov/genbank/.

In vitro Cultivation of the Feline Adenovirus Isolate

Pharyngeal and rectal swabs were taken from a cat seropositive and PCR positive for AdV (Lakatos et al., 1999), and immersed in 2 mL RPMI-1640 with 10% heat-inactivated fetal bovine serum (FBS) and 50 μg/mL gentamicin (Sigma, St. Louis, MO, United States) at 4°C. No adenovirus was being grown in the laboratory where and when the feline virus was isolated. After filtering through an 0.22 μm disposable filter (Millipore, Bedford, MA, United States), this inoculum was added to a culture of HeLa 90% confluent cells for 1 h at 37°C, and then the culture was adjusted to 5 mL with DMEM, 2% FBS. After 7–14 days detached cell suspensions were transferred to fresh flasks for three blind passages. One of the triplicate cultures developed cytopathic effect (CPE) after the third blind passage, and part of that culture's supernatant was filtered through a 0.45 μm filter. These filtrates were used to infect fresh HeLa and VERO cultures resulting in the appearance of the same CPE. Next, several human and animal cells lines were tested for adenovirus infection using a low multiplicity of infection (moi) to follow the gradual development of cytopathic effect due to the spread of the virus in consecutive generations.

[1] http://hadvwg.gmu.edu

Mammalian epithelial and fibroblast cultures were grown in DMEM, immune cells were maintained in RPMI-1640 with 10 mM HEPES, both medium completed with 10% FBS and 40 mikrog/ml gentamicin, while 2% or 7% FBS (Sigma, St. Louis, MO, United States), respectively, were added postinfection. Clarified supernatants of HeLa, CRFK, and M426 cells were used as virus stocks for subsequent infection of other cultured cells. For infection, 1 ml virus stock containing 2–4 × 10^4 $TCID_{50}$ (equals to 0.007–0.015 moi.) was used in a 25 cm^2 tissue culture flask. CPE was scored up to 14 days (**Tables 1**, **2**). Permissivity was assayed by immunofluorescence and immunocytochemistry assays (IFA and ICH, see below). Mock infected cells were adenovirus negative through the isolation process and permissivity studies tested by the same assays. Virus stocks were stored at −20 and −70°C.

Purification and Concentration of the Feline Adenovirus Isolate

HeLa cells were cultured in T75 tissue culture flasks until they reached 90% confluency. After infection with feline adenovirus, cells were cultured until cytopathic effect became visible as rounding of cells. Before the cells detached from the surface they were washed into the medium by pipetting, and then subsequently centrifuging at 1,000 *g* for 5 min. The supernatant was decanted, and cells were resuspended in the remaining supernatant. Viruses were released from the cells by three freeze/thaw cycles. Viruses were harvested by centrifugation at 3,000 *g* for 10 min then the supernatant was collected and clarified through a 0,45 μm filter. Feline adenovirus isolate was purified and concentrated with the purification filter

TABLE 1 | Replication of the feline adenovirus in different animal cell lines.

Species	Origin of cells	Cell line	Cytopathic effect	Viral antigen
Chinese hamster (*Cricetulus griseus*)	Ovarian epithelial	CHO-K1 (ECACC 85050302)	No CPE	IFA−, IHC−
Mouse (*Mus musculus*)	Fibroblast	3T3-L1 (ATCC CL-173)	No CPE	IFA−, IHC−
Rabbit (*Oryctolagus cuniculus*)	Kidney epithelial	RK-13 (ATCC CCL-37)	CPE+	IHC+
Cat (*Felis catus*)	Kidney epithelial	CRFK (ATCC CCL-94)	CPE+	IFA+, IHC+
Dog (*Canis familiaris*)	Kidney epithelial	MDCK (ATCC CCL-34)	CPE+	IHC+
Pig (*Sus scrofa*)	Kidney epithelial	PD-5 (ECACC 93120830)	CPE+	IFA+
	Kidney epithelial	PK-15 (ATCC CCL-33)	CPE+	IFA+
Rhesus macaque (*Macaca mulatta*)	Fibroblast (lung)	DBS-FRhL-2 (ATCC CL-160)	CPE+	IFA+
Green monkey (*Cercopithecus aethiops*)	Kidney epithelial	VERO (ATCC CCL-81)	CPE+	IFA+
	Kidney epithelial	COS7 (ATCC CRL-1651)	CPE+	IFA+, IHC+
Owl monkey (*Aotus trivirgatus*)	Kidney epithelial	OMK (ATCC CRL-1556)	CPE+	IFA+

ATCC, American type culture collection; ECACC, European collection of cell cultures; CPE, cytopathic effect; IFA, immunofluorescent antibody; IHC, immunohistochemistry.

TABLE 2 | Replication of the feline adenovirus in different human (*Homo sapiens*) cell lines.

Type of cells	Origin of cells	Cell line	Cytopathic effect	Viral antigen
Fibroblasts	Lung fibroblast	M426 (ATCC PTA-5244)	CPE+	IFA+
	Primary human foreskin fibroblasts	–	CPE+	IFA+
Epithelial cells	Cervical cancer	HeLa (ATCC CCL-2)	CPE+	IFA+
	Embryonal kidney	293 (ATCC CRL-1573)	CPE+	IFA+
	Breast adenocarcinoma	MCF7 (ATCC HTB-22)	CPE+	IFA+
	Melanoma	MEWO (ATCC HTB-65)	*CPE−/+	*IFA−/+
Neural cells	Glioblastoma-astrocytoma	U87 (ATCC HTB-14)	CPE+	IFA+
	Glioblastoma	A-172 (ATCC CRL-1620)	CPE+	IFA+
Leukocytes	Human cord blood mononuclear cells (primary)	–	Degeneration, No CPE	IFA+
	Adult human peripheral blood lymphocytes (primary)	–	Degeneration, No CPE	IFA+
B-lymphocytes	Mature B-cell line	Bjab (DSMZ ACC 757)	No CPE	IFA+
	EBV transformed B-cell line	LCL- 8664 (ATCC CRL-1085)	No CPE	IFA+
	Human histiocyte lymphoma	U937 (ATCC CRL-1593.2)	No CPE	IFA+
T-lymphocytes	Acute lymphoblastic leukemia	MOLT-3 (ATCC CRL-1552)	No CPE	IFA+
	Acute lymphoblastic leukemia	HSB-2 (ATCC CCL-120.1)	No CPE	IFA+
	Lymphoblastic lymphoma	Sup-T1 (ATCC CRL-1942)	No CPE	IFA+
	Acute T-cell leukemia	E6.1 (ATCC TIB-152)	No CPE	IFA+

*ATCC, American type culture collection; ECACC, European collection of cell cultures; DSMZ, Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH/German collection of microorganisms and cell cultures; CPE, cytopathic effect; IFA, immunofluorescent antibody; IHC, immunohistochemistry. *MEWO, intracellular virus antigens and cytopathic effect developed from 7 days Postinfection.*

(ViraBind^{TM} Adenovirus Purification Kit, Cell Biolabs, Inc., San Diego, CA, United States) subsequently eluted with the elution Buffer (25 mM TRIS, pH 7.5, 2,5 mM Mg_2Cl, 1 M NaCl). For storage, 10% sterile glycerol was added. The final concentration of the virus was 1.35×10^7 infectious unit/ml as titered on HEK 293 cells. Cells were infected with 10-fold serial dilution of feline adenovirus. At 48 h postinfection, before CPE becomes visible, intracellular hexon antigens were detected by immunocytochemistry (see below). Based on the average number of infected cells per microscopic field and the dilution factor, the viral titer was calculated as infectious unit/ml according to the following formula.

$$\text{Infectious unit/ml (ifu/ml)} = \frac{\begin{array}{l}\text{(Average positive cells/field)} \times \\ \text{(number of fields/well)} \times \\ \text{(dilution factor)}\end{array}}{0.1\ \text{ml}}$$

The data were expressed as mean ± standard deviation (SD).

Immunofluorescence Assays (IFA) to Detect Virus Antigens and Antiviral Antibodies

Infected and uninfected control cells of various cell lines (**Tables 1**, **2**) were washed in phosphate buffered saline (PBS) and spotted on Teflon coated slides, air-dried, fixed in cold methanol-acetone and incubated with 10 μL of Adenovirus monoclonal antibody (Bartels VRK, Intracel, Issaquah, WA, United States), washed and stained with FITC conjugated anti-mouse IgG with and Evans blue counterstain as the secondary antibody. The cells were screened for specific immunofluorescence with a UV microscope. Human embryonic kidney 293 (HEK 293) cells infected with HAdV-2 and HAdV-7 were used as positive controls. Mock-infected cells and an irrelevant antibody to human herpesvirus 7 (RK4, Advanced Biotechnologies Inc., Columbia, MD, United States) served as negative controls. Furthermore, sera obtained from several male and female random domestic cats brought to the Selbyville Animal Hospital in Delaware United States were tested at a 1:20 dilution by IFA using the highly permissive M426 human cell line infected with FeAdV to assess the distribution of IgG antibody. Antibody-positive feline sera were titered.

Immunocytochemistry Staining (ICH)

The presence of intracellular adenovirus hexon antigen in selected cell cultures was also visualized by immunoassay (Quick Titer Adenovirus Titer Immunoassay Kit, Cell Biolabs, Inc., San Diego, CA, United States). Cells at a concentration of 4×10^4 cells/well were seeded in a 96-well tissue culture plate and, at 90% confluency, they were infected with the purified feline adenovirus at 5 moi. On day 3, when CPE was detected, the culture medium was decanted. Cells were fixed by methanol at −20°C for 20 min; subsequently washed with phosphate buffered saline (PBS) three times for 5 min each then blocked with 1% bovine serum albumin (BSA) in PBS for 1 h at room temperature on an orbital shaker. After removing the blocking solution, cells were incubated with the anti-hexon antibody for 1 h at room temperature on an orbital shaker. After washings (see above) cells were incubated with the horseradish peroxidase (HRP)-conjugated secondary antibody at room temperature for 1 h on an orbital shaker followed by washing five times. Wells were stained with freshly diluted 3,3′-diaminobenzidine (DAB) working solution for 10 min to visualize infected cells. Finally, wells were washed with PBS twice. Cells were examined in a light microscope. Infected cells were brown; the non-infected were not stained. The same negative and positive controls were used as in IFA studies.

Electron Microscopy (EM)

After infection, when maximal CPE was achieved, CRFK, HeLa, PD-5, and M426 cells (4–5 days post infection) were centrifuged and fixed in glutaraldehyde-paraformaldehyde. Thin sections were stained. Supernatant samples were processed for negative staining. Both infected and uninfected samples were studied electron microscopically as described (Ongrádi et al., 2000).

Polymerase Chain Reaction (PCR), Sequencing of the Hexon and Fiber Genes and Aligned With Published Sequences

Hexon specific PCR was carried out for subgenus and type determinations in Hanover, Germany in 1999. FeAdV was propagated several times in HeLa cultures. Viral DNA was prepared. The sample was tested for species specificity in PCR using a series of primers HsgA1 to HsgF2. DNA of serotypes belonging to the corresponding species were controls as described (Pring-Åkerblom et al., 1999). Since the feline sample showed positivity with species C primers, studies were extended to specific primers for HAdV-1, 2, and 5 (Pring-Åkerblom et al., 1997a). Next, the loop 1_4 region and parts of the flanking conserved hexon gene regions were amplified with $H1_4$/$H1_42$ primer pairs (Pring-Åkerblom et al., 1999). Primer pairs specific for HAdV-1 knob region were used to detect the fiber gene of the feline isolate as published (Pring-Åkerblom and Adrian, 1995a). PCR products of the hexon loop 1_4 region were cloned into pUC18. Similarly, the complete hexon and fiber genes were sequenced using specific internal primers (Pring-Åkerblom and Adrian, 1993; Pring-Åkerblom et al., 1995, 1997b). DNA sequencing of an European HAdV-1 isolate [Netherlands, 1970, strain code 1038, genotype D11 (cited as D11)] (Adrian et al., 1990; Pring-Åkerblom et al., 1999) was carried out in the same way, and sequence alignment was done using DNASIS (Pharmacia, Uppsala, Sweden) as described (Pring-Åkerblom and Adrian, 1995b; Pring-Åkerblom et al., 1999). Recently, fiber and hexon sequences have been aligned with counterparts published in GenBank (Figures 10, 11) using Multiple Sequence Comparison by Log-Expectation [MUSCLE] (2018) with parameters provided in MUSCLE. The MEGA 7.0 (Molecular Evolutionary Genetic Analysis) software was used to generate a phylogenetic neighbor-joining tree applying the Neighbor-joining method using a Kimura-2-parameter model for calculating evolutionary distance, and 1,000 replicates for bootstrap analysis. Only those sequences could be matched that have corresponding regions in FeAdV, because in the case of other isolates, slightly longer or shorter fragments were available. The G + C ratio of available sequences was calculated by a formula (G + C)/(A + T + G + C) × 100, and compared to the homologous genes of known isolates.

RESULTS

Cell Biology

Some "toxic" effect was seen at the edge of the HeLa monolayers for a few days post inoculation with the rectal specimen. During the third passage of this culture, the cytopathic effect (CPE) developed. The cells became refractile, rounded up, and formed grape-like clusters typical of species C type AdV. Transferring these cells or their cell-free supernatants (either filtered or unfiltered) to fresh HeLa cultures elicited the same CPE in 3–4 days (**Figure 1**). The virus titers from the supernatants of the HeLa cells peaked at 4 × 10^4, 50% tissue culture infectious dose ($TCID_{50}$) on day 5. This result shows that an infectious agent is obtained from the feces of the cat.

Meanwhile, another test showed that the infectious agent is an adenovirus. They have a very narrow host range. It was logical to establish the host range of the isolate, at least *in vitro*. Several human, feline and other mammalian cells lines proved to be permissive for the feline virus isolate, although there were some differences in timing and presentation of CPE (**Tables 1**, **2**). On day 5, the virus titer produced by VERO cells reached its maximal level at 2 × 10^4 $TCID_{50}$. Human anti-AdV antibodies in immunofluorescent assays showed strong fluorescence in both nucleus and cytoplasm, as well as in the cytoplasmic membrane. The course of infection, peak titers, and distribution of intracellular virus antigens were similar in M426, HEK 293, OMK, rhesus monkey fibroblasts, CRFK and PD-5 and other mammalian cells to those found in HeLa and VERO cultures.

In U-87 cells, a neural cell line, CPE appeared on day 5 and reached 100% maximum on day 7, while in A-172 cells, another neural cell line, it became apparent on day 7 and reached maximal level at 75% in 10 days. In MEWO cells, the first sign of CPE was observed on day 7 only; it progressed very slowly as compared to the CPE seen on HeLa and CRFK cultures. Maximal CPE developed on day 10 (**Figure 2**). To compare adenovirus replication in HeLa, CRFK, and MEWO cells, 0.01 and 1.0 moi inocula were used to infect cultures, followed by virus titration during replication. FeAdV replication in CRFK cells depended on the input virus: 0.01 moi hardly resulted in virus replication, while 1.0 moi elicited FeAdV replication as high as in HeLa cultures. MEWO cells could be regarded as semi-permissive due to extremely slow adenovirus replication (**Figure 3**). No CPE were detected in CHO and 3T3-L1 cultures (**Figure 2**).

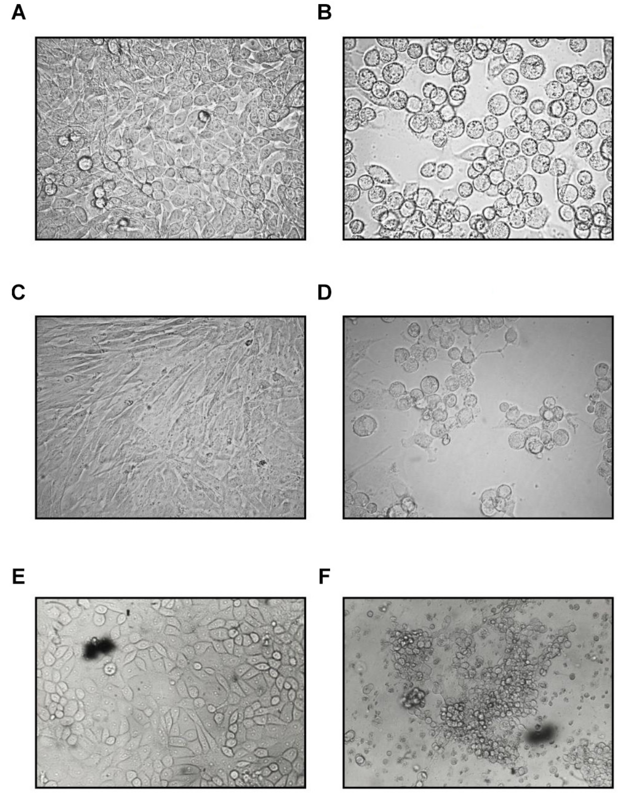


FIGURE 1 | Cytopathic effect of the feline adenovirus isolate in permissive cells as compared to uninfected cells on day 5 postinfection (400×). Uninfected **(A)** and infected **(B)** HeLa cells, uninfected **(C)** and infected CRFK **(D)** cells, uninfected **(E)**, and infected **(F)** PD-5 cells. Infected cells form grape-like clusters.

Human adult peripheral blood lymphocytes and cord blood lymphocytes did not exhibit CPE, but the degeneration of infected cultures was observed, and the IFA staining was restricted to the cytoplasm. No CPE and weak cytoplasmic fluorescence was detected in the human T lymphoid E6.1, HSB-2, MOLT-3, Sup-T1, B lymphoid Bjab and LCL and monocytic U-937 cultures, suggesting that these unstimulated cells may be non-permissive for FeAdV replication (**Table 2**).

Immunofluorescence, Immunohistochemistry, and Electron Microscopy

Verification of the infectious agent occurred partially by classical antigen detection. Using anti-adenoviral antibodies, immunofluorescence and immunocytochemistry resulted in identical positive or negative results (**Tables 1**, **2**). Adenovirus antigens were detected in all mammalian cells tested (**Figure 4**), except in CHO-K1 and 3T3-L1 cultures. Mock-infected cells remained negative. An irrelevant monoclonal antibody to human herpes virus 7 (RK4, 1:50 dilution) resulted in no fluorescence in cells with or without CPE.

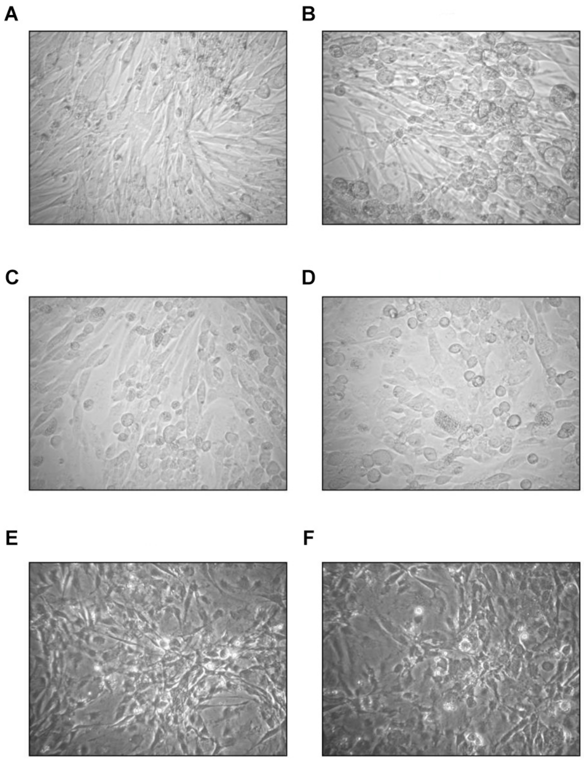


FIGURE 2 | Effect of the feline adenovirus isolate in semi- and non-permissive cells as compared to uninfected cultures (400×). On day 7, uninfected **(A)** and infected **(B)** MEWO cells, on day 3, uninfected **(C)** and infected **(D)** CHO-K1 cells, on day 4 using phase contrast illumination, uninfected **(E)** and infected **(F)** 3T3-L1 cells.

Supernatants of infected HeLa and M426 cells stored at −20°C for 3 years, and other supernatants of HeLa and CRFK cells stored at −20 or −70°C for 13 years resulted in the same level of infectivity as fresh virus stocks did, which demonstrates that the virus is very stable. Supernatants of infected PD-5 cells contained a large number of 88 ± 4 nm icosahedral particles resembling adenoviruses, but their fibers could not be visualized (**Figure 5**). The cytoplasm of infected PD-5, M426, HeLa, CRFK cells contained identical paracrystal arrays characteristic for adenoviruses (**Figure 6**). Electron microscopy showed that a single infectious agent was isolated and only this one infected and replicated in cell cultures.

Fifteen domestic feline sera from Delaware were tested by IFA on M426 cells infected with FeAdV isolate. Of these only three were completely negative. The remaining 12 showed staining varying from only nuclear to whole cell fluorescence; three sera had titers of 1:640 or greater. No IFA signal was seen when these sera were used to stain M426 cells that were not infected with FeAdV isolate. Any specific binding between feline antibodies and feline cellular antigens was avoided in IFA tests by using human M426 cells.

Sequencing and Alignment of Hexon and Fiber Genes

The hexon and fiber genes of the AdV isolate were sequenced due to their importance in adenovirus biology and classification. Species and type-specific PCR using species C and type HAdV-1 specific primers resulted in two amplimers one at bp 269 and the second at bp1049, respectively. Negative results were obtained using other primers. This indicates that the FeAdV isolate is closely related to HAdV-1. The hexon gene of the feline isolate was sequenced and compared to that of an HAdV-1 isolate [EMBL accession number X67709/26 (Pring-Åkerblom and Adrian, 1993)]. The mapped 2890 base pair region showed 99% identity with HAdV-1 (**Supplementary Figure S1**). Three nucleotide differences (bases 378, 1890, and 2196) in the feline sequence were in the last place of a triplet code and resulted in the same amino acids (aa) as are found in the human protein (**Supplementary Figure S2**). Sequence alignment with other published sequences shows that the hexon gene of the FeAdV isolate is most closely related to HAdV-1 (**Figure 6**) isolated in Marseille, France (Cassir et al., 2014), although its published sequence is 171 nt shorter at 3′ end and 341 nt at 5′ end. Two nucleotide differences were found. One was shown at nt 316 where C was replaced by T in the isolate from France, at nt 1,476 G was replaced by A in the same isolate. The homology of the comparable regions is 99.85%. The G + C content was higher (50.02%) in FeAdV isolate than in the isolate from Marseille (45.61%). None of the nucleotide alterations resulted in amino acid changes; leucine was coded at both sites. This phylogenetic tree based on hexon homology suggests that FeAdV isolate is related to several HAdV-1 isolates from Germany, Japan, and Taiwan (**Figure 6**). The sequence of the feline fiber gene was compared to the genotype of the D11 human counterpart (**Supplementary Figure S3**); it consisted of 1749 nucleotides that code for 582 amino acids (**Supplementary Figure S4**). There were a total of 12 differences in sequence found (**Table 3**), none in the tail (aa 1–44); three in the shaft (aa 45–401), all of which result in a change in amino acid, two with a change in charge; and nine in the knob region (aa 402–582). Of the alterations in the knob region, five were neutral for the amino acid sequence, and only one alteration resulted in a charge difference. One difference was in the conserved area of the knob; namely, amino acid 442 went from an arginine to a lysine. Differences resulting in an altered amino acid in the fiber protein affected the first nucleotide of the triplet code in four cases (base 220, 1015, 1240, and 1414), the middle nucleotide in two cases (base 596 and 1325), and the last nucleotide in one case (base 1581). Comparison of the fiber gene of FeAdV to that of other isolates shows that the feline isolate is most closely related to the first adenovirus isolated in 1953 (GenBank accession No AB125750, Adhikary et al., 2004a GenBank Accession No AB108423, Adhikary et al., 2004b, referring to adenoid71 by Rowe et al., 1953). Only one nucleotide alteration was found at position 981 in the shaft region without amino acid change; this is the last unit of that triplet code. Nucleotide homology between FeAdV isolate and adenoid 71 is 99.83%. The G + C content of the feline fiber is insignificantly higher (44.71%) than that of the isolate

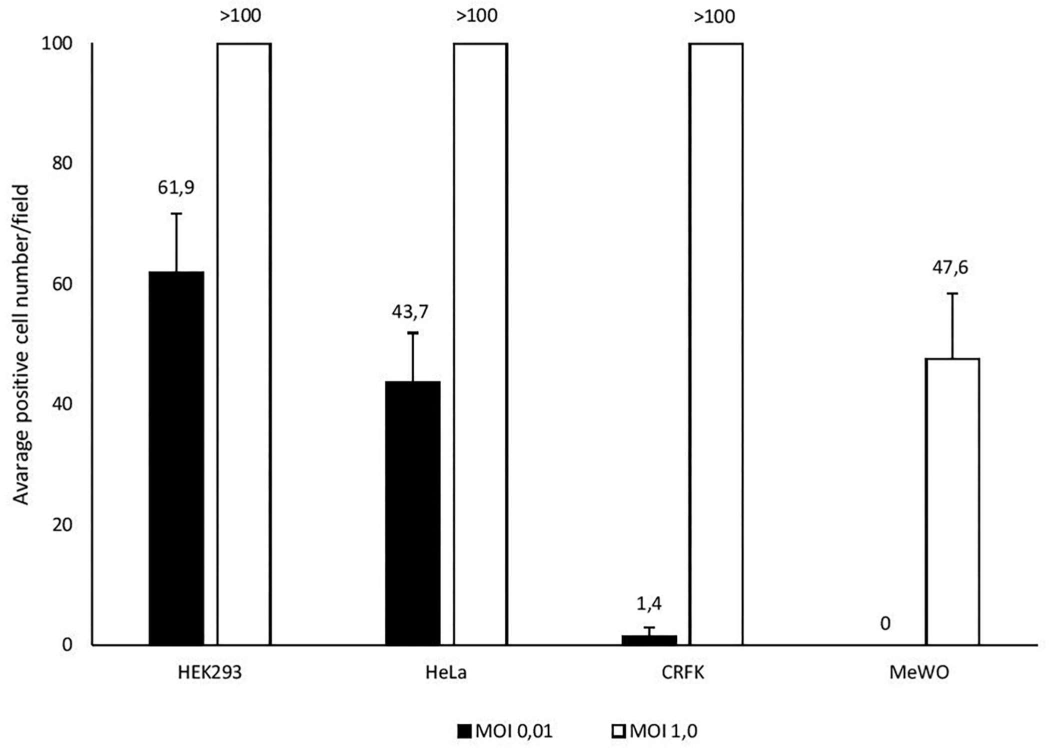


FIGURE 3 | Different susceptibility of cell lines (HEK293, HeLa, CRFK, and MEWO) to feline adenovirus infection detected by immunocytochemistry. Cells were infected with different multiplicity of infection and positively stained cells were counted in a microscopic field 48 h post infection.

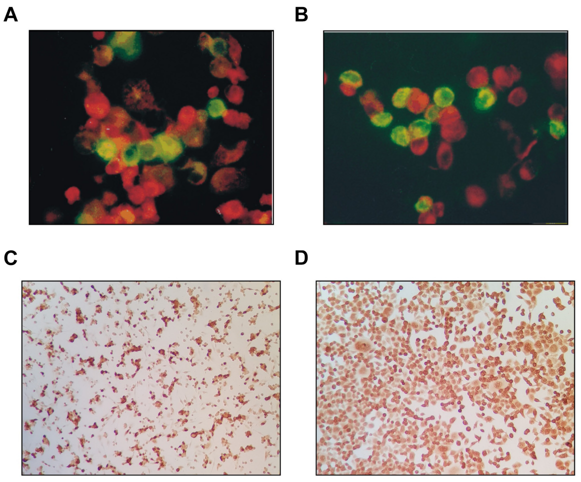


FIGURE 4 | Adenovirus antigens detected in both the nucleus and cytoplasm of CRFK cat kidney cells **(A)** and HEK293 (human embryonic kidney) **(B)** by FITC-conjugated anti-human adenovirus antibody as compared to uninfected cells counterstained red (400×). The same antigens were detected in CRFK **(C)** and HeLa **(D)** cells by immunocytochemistry (dark cells).

from Germany or Rowe's isolate (44.28%). Interestingly enough, isolates from Taiwan are relatively closely related to FeAdV fiber gene (**Figure 7**).

DISCUSSION

Eighty percent of feline sera tested for antibodies were positive suggesting a very high prevalence of this strain of FeAdV in United States domestic cats. High percentage of seropositive sera using the FeAdV as antigen in comparison with 9.8–26% positivity using HAdV-1 as antigen (Lakatos et al., 1996, 2000) clearly shows that, the structure of antigens in the two AdVs is different, and the feline antibodies possess a higher avidity to the antigens of the feline AdV. This is in good correlation with the different amino acid sequence and consequent antigenicity between the human and feline AdV isolates. Following these pilot studies, more sera need to be tested especially those from cats showing some clinical manifestations. A limitation through the course of feline adenovirus serological studies is that human hexon antigen was used for testing adenovirus serology in European and American cats (Lakatos et al., 1996) by ELISA, but FeAdV antigen was used to test serology in the recent pilot study by IFA. Although the results unambiguously suggest important avidity differences, using both methods for testing all sera could be ideal in the future.

HeLa cells were used for isolation due to their high sensitivity for AdVs and, due to the lack of information on feline adenovirus cultures, reagents, diagnostic tools. The difference in the permissiveness for FeAdV of the various cell lines tested may be due to characteristics of the cells themselves. Identical CPE,

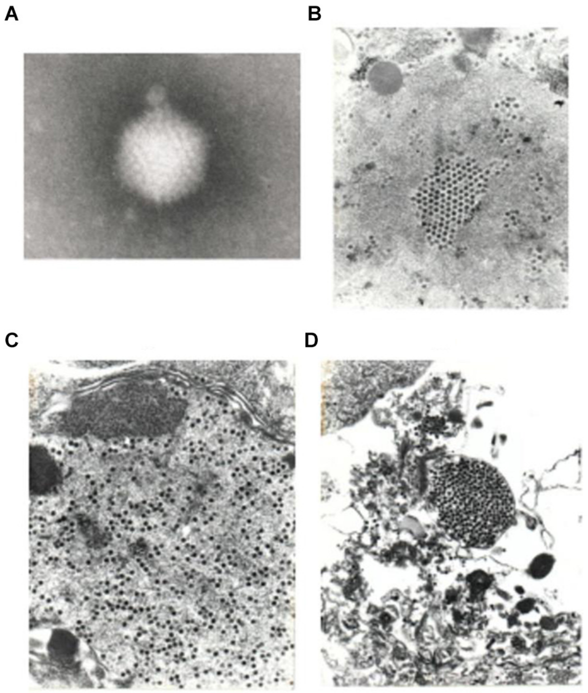


FIGURE 5 | Electron micrograph of a negatively stained virus particle obtained from the supernatant of infected PD-5 culture **(A)**. The icosahedral shape and size (88 ± 4 nm) is characteristic for adenoviruses. Fibers could not be visualized (80,000× plus picture magnification). A pseudo array formed by adenoviruses in the nucleus of a PD-5 cell **(B)**, viruses scattered in the cytoplasm **(C,D)** (10,000× plus picture magnifications).

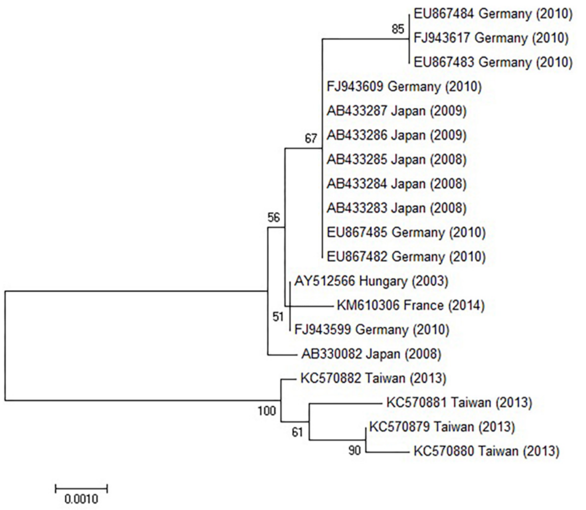


FIGURE 6 | Phylogenetic tree constructed using HAdV-1 hexon genes retrieved from GenBank. The bar indicates the number of substitutions per site.

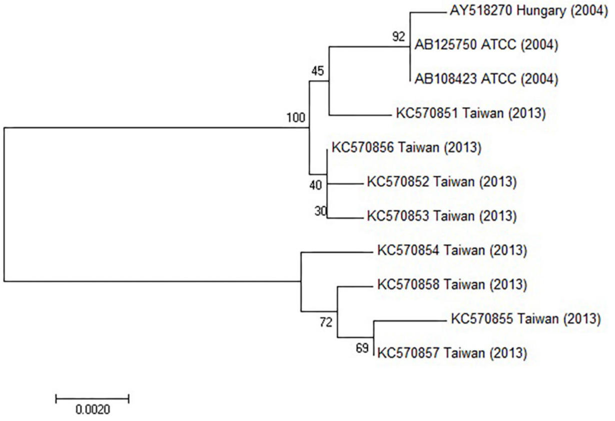


FIGURE 7 | Phylogenetic tree constructed using HAdV fiber gene sequences retrieved from GenBank. The bar indicates the number of substitutions per site.

IFA, and IHC positivity suggest that this feline adenovirus isolate can infect and replicate in several cell lines, that express both Coxsackie-adenovirus receptor (CAR, Howitt et al., 2003) and $\alpha_v\beta_3$ or $\alpha_v\beta_5$ integrins as coreceptors (Wickham et al., 1993). MDCK cells express CAR in the tight junctions, but FeAdV was able to infect them possibly by opening tight junctions (Sharma et al., 2012). CHO (Coyne and Bergelson, 2005) and 3T3-L1 cells (Orlicky and Schaack, 2001) cannot be infected by FeAdV, due to lack of CAR expression. The difference in the permissivity of CRFK cells using high inocula and low permissivity using small inocula might be an important finding in our study. Infection of CRFK cells with fowl adenovirus using relatively small inoculum (moi 1) resulted in binding of adenovirus particles through their CAR-dependent long fiber, but no virus replication was detected (Taharaguchi et al., 2012). CAR-dependent adenovirus should bind to integrins $\alpha_v\beta_3$ or $\alpha_v\beta_5$ for infection and replication on the surface of CRFK cells (Mahalingam et al., 2014). One can presume that in our experiments, a small amount of FeAdV inoculum either could not open cellular tight junctions for an autocatalytic availability of CAR molecules (Sharma et al., 2012) or could not activate integrin molecules from their inactive state (Mahalingam et al., 2014). Semi-permissivity of MEWO cells might be due to low expression of integrins that take part in AdV internalization (Menter et al., 1995), although these cells were successfully infected by HAdV type 5 vectors (Oka et al., 2002; Sarkar et al., 2008). In good correlation with our results, it is known that species C of HAdV has a low affinity for neural cells because the primary cell receptor for the virus, CAR expression is highly variable on neural cells (Skog et al., 2002).

The IFA and ICH tests detected FeAdV antigens in human leukocytes, B and T lymphocyte cultures suggesting that the feline adenovirus can infect these cells. On the contrary, CPE was not observed in the same immune cells meaning that, no significant virus replication took place in these cultures. The observed phenomenon is in good correlation with earlier publications, namely that, the serotypes of species C are known to establish lifelong persistence in lymphoid tissues and, cellular stimuli are required to reactivate viruses, and elicit virus replication (Lauer et al., 2004).

TABLE 3 | Difference in base sequence and resulting amino acid sequence of feline adenovirus fiber protein as compared to analogous human adenovirus (HAdV) type 1 nucleotides and proteins.

Base #	Feline base	D11 human base	ATCC AB125750 base	Amino acid #	Location	Feline amino acid (charge)	D11 human amino acid (charge)	ATCC 125750 amino acid
Shaft portion (amino acids 45–401)								
220	A	G	A	74	*2. pseudorepeat*	Lysine (+)	Glutamic acid (−)	*Lysine (+)*
596	A	G	A	199	*9. pseudorepeat*	Asparagine	Serine	*Asparagine*
981 (1047*)	A	A	G	327	*17. pseudorepeat*	*Lysine (+)*	*Lysine (+)*	*Lysine (+)*
1015	C	A	C	339	*18. pseudorepeat*	Histidine (+)	Asparagine	*Histidine (+)*
Knob portion (amino acids 402–582)								
1240	C	T	C	414	AB loop	Histidine (+)	Tyrosine	*Histidine (+)*
1248	*C*	*T*	*C*	*428*	*BC loop*	*Cysteine*	*Cysteine*	*Cysteine*
1325	G	A	G	442	CD loop	Arginine (+)	Lysine (+)	*Arginine (+)*
1341	*T*	*C*	*T*	*447*	*CD loop*	*Proline*	*Proline*	*Proline*
1395	*A*	*G*	*A*	*465*	*DE loop*	*Glycine*	*Glycine*	*Glycine*
1414	A	G	A	472	DE loop	Serine	Glycine	*Serine*
1581	A	T	A	527	GH loop	Glutamic acid (−)	Aspartic acid (−)	*Glutamic acid (−)*
1659	*C*	*A*	*A*	*553*	*HI loop*	*Serine*	*Serine*	*Serine*
1743	*G*	*A*	*A*	*581*	*Carboxy terminal*	*Glutamine*	*Glutamine*	*Glutamine*

**Nucleotide number of AB125750 published in ATCC. Silent amino acid alterations as compared to the feline amino acids are shown in italic.*

Surprisingly and uniquely, FeAdV replicated well in several mammalian cell lines demonstrating a wide cell tropism. As the alterations in the hexon gene did not result in any change in amino acid sequence, it is unlikely that this protein showed any functional change. For explaining this unusual phenomenon, it is speculated that the amino acid differences from the human sequence in the fiber might account for the expanded tropism. The fiber is the major determinant of tropism, the globular knob region of the fiber polypeptide attaches to CAR (Pring-Åkerblom and Adrian, 1995a,b; Eiz and Pring-Åkerblom, 1997; Roelvink et al., 1999). Each knob molecule contains b-strands (A to J) connected by prominent loops. Several conserved amino acid residues on the AB loop, B b-strand, CD-, DE-, FG-, and HI loops are required for binding to CAR (Roelvink et al., 1999), although contact residues are not well conserved (Howitt et al., 2003). S408 and Y477 crucial binding residues are found in both HAdV-2 (Howitt et al., 2003) and our HAdV-1 D11 and FeAdV isolates. Although four amino acid differences from the human isolate affected the AB-, CD-, DE-, and GH loops (**Table 2**), only the change of amino acid 442 from lysine to arginine is likely to have any functional significance that might result in a change of tropism or pathogenicity (Pring-Åkerblom and Adrian, 1994, 1995b; Eiz and Pring-Åkerblom, 1997), because it is the only alteration in the conserved CAR binding area of the CD loop (**Figure 7**). Altered arginine 442 and glutamic acid 527 is identical to species E HAdV-4 having an arginine in the same position (Pring-Åkerblom and Adrian, 1995a,b; Pring-Åkerblom et al., 1998). Binding FeAdV to porcine cells line raises the possibility that its fiber knob possesses structures similar to the porcine (P) AdV-4 and PAdV-3 (Nagy et al., 2002). Among the CAR binding conserved sequences common in HAdV-1 and FeAdV, only four amino acids in the AB loop, two in the FG loop of PAdV-4 (Kleiboeker, 1995), three amino acids in PAdV-3 (Reddy et al., 1995), two amino acids of PAdV-5 AB or FG loops (Nagy et al., 2002) are identical. None of the altered amino acids of FeAdV fiber knob are found in any of the PAdV. These data exclude a direct relationship between FeAdV and PAdV. The structure of feline CAR is not yet known. CAR is highly conserved; the domain of human and porcine CAR that is known to interact with the fiber shows 93.7% homology (Griesche et al., 2008).

No differences in the amino acid structure were detected through alignment of FeAdV fiber with adenoid 71 virus or alignment of FeAdV hexon with the most related isolate from France. This suggests that FeAdV isolate might be relatively distant from the European HAdV-1 isolate D11, but could be phylogenetically nearer to American adenoid 71 (Rowe et al., 1953) and Asian isolates. Detection of FeAdV in a human sample in Japan also suggests this relationship (Phan et al., 2006). Higher G + C content in the hexon gene of FeAdV isolate than in the isolate from Marseille suggests that FeAdV hexon gene possesses increased genomic stability, although drawing conclusions from the limited span of the genome seems to be very early.

Two of the alterations in the fiber shaft resulted in gaining two excess basic amino acids in the 2nd and 8th pseudo repeat 15-amino acid motifs. It is presumed that this might cause a functional change allowing more flexibility of the rod-like shaft, thus allowing it to adapt to a wider variety in cell types as it has been known with other C types (Lion, 2014). Particle binding to CAR is followed by adjusting the arginine-glycine-aspartic acid (RGD) motif of the penton base to cell membrane integrins for subsequent internalization (Fechner et al., 1999). In recombinant AdV vectors, inserting an RGD motif into the HI loop results in CAR-independent cellular entry via integrins (Madisch et al., 2007). Insertion of RGD into the feline parvovirus also extends its tropism to human tumor cells (Maxwell et al., 2001). The fiber knob of PAdV-4 contains an RGD motif beginning at residue 361 and surrounded by upstream sequences rich in alanine and glutamic acid (Kleiboeker, 1995). This structure resembles the penton base of species C HAdV, in which the RGD motif with downstream alanine and glutamic acid-rich sequences mediate integrin binding (Fechner et al., 1999). In the fiber knob of FeAdV the altered amino acid 442 forms an arginine-glycine-serine (RGS) motif at the beginning of the CD loop which is surrounded by alanine residues at positions 438, 440, 446, and 445 and a glutamic acid at 463. Furthermore, two residues downstream the RGS motif in FeAdV and the RGD motif in PAdV-4 are identical: leucine-alanine (LA). These speculations raise the possibility that the feline RGSLA moiety might mediate an alternative virus internalization into cells with low integrin expression, as this could be the way of infection of MEWO cells.

Isolation of FeAdV suggests that an AdV can infect cats. A concern is whether this isolate is a cross-contaminant or really obtained from an animal in which virus replication occurred. The history of the cat yielded the FeAdV isolate suggests; it is not a contaminant. Cited as: "From a 2-year-old domestic cat kept as a single pet in isolation, pharyngeal and rectal swab samples were taken twice at a 12 months interval. At the beginning of the examinations, the cat suffered from transient hepatic failure. Subsequently, the animal was repeatedly examined by a group specific indirect ELISA test, and found highly seropositive for adenovirus hexon antigens throughout 18 months. . . . The first rectal and the rectal and pharyngeal samples taken 1 year later gave amplification bands of identical size with the positive control. . . The positive PCR results are suggestive of persistent infection and shedding of adenovirus in the examined animal. . . The DNA sequence of the three PCR products (GenBank Accession number AF172246 are identical)" (Lakatos et al., 1999). The first rectal sample was used for virus isolation. The 301 bp fragment of the hexon gene product obtained directly from the feces is identical to the sequence amplified from the virus isolate (GenBank Accession number AY512566) (**Supplementary Figure S1**).

The genome of FeAdV isolate spanning between the fiber and hexon genes and the first half of the genome has not been characterized; theoretically, they could carry several mutations, or be recombinants, as novel HAdV types have shown such structures (refs in Lion, 2014). Non-human related AdV types also might establish an opportunistic infection in immunocompromised cats, as sequencing of the small fragments of hexon and polymerase genes in an archived sample (Kennedy, 1993) has recently shown (Lakatos et al., 2017), but whole genome sequencing has not been done. Serological studies and viral DNA detection, isolation of a replication competent adenovirus from a cat and, its basic characterization are a milestone in the series of feline adenovirus studies. Lack of the whole genome sequencing can be seen as a limitation of both adenovirus studies. As a next milestone, complete sequence analysis could yield more information on the origin of feline adenoviruses.

Interspecies transmission and recombination of adenoviruses are not a curiosity, e.g., it has been documented in the case of HAdV-4 (Dehghan et al., 2013), for other cases in species Human mastadenovirus C see Virus Taxonomy 2018b Release (2018). The broad tissue permissivity and the few differences between the feline and human viral polypeptides also strongly suggests an anthropozoonosis from human to cat or vice versa. HAdV-1 is widespread all over the world. Consequently one of the possibilities is that the FeAdV isolate is a natural variant of HAdV-1 adapted to cats. A case report from Japan found the same AdV cluster in the fecal specimen collected from a 1-year-old girl with acute gastroenteritis. They shared 100 and 97% identities at the amino acid levels of hexon and fiber genes, respectively, with corresponding FeAdV genes. The virus was designated as 6277JP, amino acid sequence was submitted to DDRJ DNA/GenBank database (accession number DQ336392) (Phan et al., 2006). A very broad phylogenetic tree containing several human, mammalian and avian adenovirus "subgenus" and species was constructed. Feline adenovirus, HAdV type 1/Adenoid 71 (Prototype), HAdV type 1/6277JP clustered. This publication was the first to show a close relationship between FeAdV and HAdV-1 (Phan et al., 2006).

Furthermore, these authors conclude that their case is proof of the interspecies transmission of HAdV-1 variant as FeAdV isolate, from a cat to the child. Lately, one isolate from 468 upper respiratory tract specimens in Brazil has shown 100% hexon gene sequence homology to the counterpart of FeAdV (Luiz et al., 2010). Both publications suggest that FeAdV is prevalent all over the world and, can infect humans, but presumably, rarely does. In a recent publication, phylogenetic analysis limited to hexon and fiber sequences has shown that the most related isolate to FeAdV was found in an intensive care unit in Marseille, France. Bronchoalveolar lavage of patients with highly immunocompromised state following lung transplantation and mechanical ventilation showed PCR positivity. Presumably, a nurse taking care of patients without wearing a mask while exhibiting respiratory signs and symptoms might have been the source of the outbreak. As no other cases of HAdVs were reported in Marseille public hospitals and among relatives of health care workers (Cassir et al., 2014), it raises the possibility of zoonotic infection from the cat of the nurse. This aspect has not been explored in the above-mentioned epidemiological survey, but it will be a must in future cases!

Detection of HAdV-1 hexon PCR positivity must be followed by sequencing to distinguish HAdV-1 and FeAdV (Jothikumar et al., 2005). Sequencing the fiber-specific amplicon also could distinguish between human and feline AdVs as recommended for other AdVs (Lion, 2014). These methods could yield exact data on the global epidemiology of FeAdV. The *in vivo* effects of the feline isolate is not known for humans, cats or *Felidae* especially, endangered big cats; these ought to be explored. Immunocompromised hosts might be at a higher risk, e.g., simultaneous FeLV (Kennedy, 1993) or FIV (Lakatos et al., 2000) infection. FeAdV has exceptional value in AIDS research: it can be utilized in both human studies and the *in vivo* feline AIDS model As speculated, the flexibility of the fiber molecule predisposes FeAdV isolate for further genetic manipulations as a gene therapy vector, especially with oncolytic potential. Depending on availability of funds, whole genome sequencing of the feline isolate is planned for such purposes.

Based on published hexon and fiber sequences (Pring-Åkerblom and Ongrádi, 1980a,b) and previous publications listed above, FeAdV isolate was included in the 14th International Committee for Virus Taxonomy Database (ICTVdb) as 00.001. Adenoviridae, 00.001.0.01.010.00.101.901 HAdV 1, feline isolate HAdV-1 approved by the Adenoviridae Study Group (Fauquet et al., 2000). The archived copy is available on the internet. Criteria of the novel taxonomy have been changed since no isolates have been included in the taxonomy[2] [Virus Taxonomy 2018b Release (2018)]. The feline adenovirus (Taxon identifier 114408) is listed in UniProt (2018) (UniProt accessed November 13, 2018). FeAdV isolate has been deposited in the American Type Culture Collection (ATCC) (AcqID-00755).

All previous and recent data taken together support the idea that cats can be infected by adenoviruses, viruses replicate in their cells, viruses are shed, an adequate immune response is mounted. We are aware of the fact that a single isolation cannot enlight the pathomechanisms of adenovirus replication in cats. Furthermore, limited information on the genetic and polypeptide structures of the single isolate cannot explain the molecular mechanism of its presumed interspecies transmission. Our considerations at the molecular level outlined here should draw attention to these particular aspects to analyze further isolates.

AUTHOR CONTRIBUTIONS

JO and DA designed the research. LC, KT, DA, PP-Å, BS, and KN conducted the research. BL and DG provided feline specimens. All authors listed had analyzed the data, made a substantial intellectual contribution to this work, and approved the manuscript for publication.

FUNDING

This work was supported in part by the start-up grant (No. 11725, Semmelweis University, Budapest, Hungary).

ACKNOWLEDGMENTS

Dr. Judith Farkas contributed to the isolation process with technical help. Electron microscopic studies by Dr. Sándor Paku (1st Department of Pathology and Experimental Cancer Research, Semmelweis University, Budapest, Hungary), Ms. Louise Robillard (Department of Microbiology, CHU Sainte-Justine Hospital, Montreal, QC, Canada), Dr. G. Williams (Advanced Biotechnologies Inc., Columbia, MD, United States) are appreciated. We are thankful to Dr. Peter Medveczky (University of South Florida, Tampa, FL, United States) for critical reading and suggestions. We are also thankful to Ms. Elaine O'Leary (HHV-6 Foundation, Santa Barbara, CA, United States) for correcting the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01430/full#supplementary-material

[2] http://ictvonline.org

REFERENCES

Adhikary, A. K., Banik, U., Numago, J., Suzuki, E., Inada, T., and Okabe, N. (2004a). Heterogeneity of the fiber sequence in subgenus C adenoviruses. *J. Clin. Pathol.* 57, 612–617. doi: 10.1136/jcp.2003.014944

Adhikary, A. K., Inada, T., Banik, U., Numaga, J., and Okabe, N. (2004b). Identification of subgenus C adenoviruses by fiber-based multiplex PCR. *J. Clin. Microbiol.* 42, 670–673. doi: 10.1128/JCM.42.2.670-673.2004

Adrian, T., Sassinek, J., and Wigand, R. (1990). Genome type analysis of 480 isolates of adenovirus types 1, 2, and 5. *Arch. Virol.* 112, 235–248. doi: 10.1007/bf01323168

Cassir, N., Hraiech, S., Nougairede, A., Zandotti, C., Fournier, P. E., and Papazian, L. (2014). Outbreak of adenovirus type 1 severe pneumonia in a French intensive care unit. September–October 2012. *EuroSurveill* 19:20914. doi: 10.2807/1560-7917.es2014.19.39.20914

Coyne, C. B., and Bergelson, J. M. (2005). CAR: a virus receptor within the tight junction. *Adv Drug Deliv. Rev.* 57, 869–882. doi: 10.1016/j.addr.2005.01.007

Dehghan, S., Seto, J., Liu, E. B., Walsh, M. P., Dyer, D. W., Chodosh, J., et al. (2013). Computational analysis of four human adenovirus type 4 genomes reveals molecular evolution through two interspecies recombination events. *Virology* 443, 197–207. doi: 10.1016/j.virol.2013.05.014

Eiz, B., and Pring-Åkerblom, P. (1997). Molecular characterization of the type-specific gamma-determinant located on the adenovirus fiber. *J. Virol.* 71, 6576–6581.

Elder, J., Lin, Y. C., Fink, E., and Grant, C. K. (2010). Feline immunodeficiency virus (FIV) as a model for study of lentivirus infections: parallels with HIV. *Curr. HIV Res.* 8, 73–80. doi: 10.2174/157016210790416389

Fauquet, C. M., Regenmortel, M. H. V., Bishop, D. H. L., Carsten, E. B., Estes, M. K., Lemon, S. M., et al. (2000). "Virus taxonomy," in *Seventh Report of the International Committee on Taxonomy Viruses*, eds M. H. V. van Regenmortel, C. M. Fauquet, and D. H. L. Bishop (Cambridge: Academic Press).

Fechner, H., Haack, A., Wang, H., Wang, X., Eizema, K., Pauschinger, M., et al. (1999). Expression of Coxsackie adenovirus receptor and alphav-integrin does not correlate with adenovector targeting in vivo indicating anatomical vector barriers. *Gene Ther.* 6, 1520–1535. doi: 10.1038/sj.gt.3301030

Gonin, P., Fournier, A., Oualikene, W., Moraillon, A., and Eloit, M. (1995). Immunization trial of cats with a replication-defective adenovirus type 5 expressing the ENV gene of FIV. *Vet. Microbiol.* 45, 393–401. doi: 10.1016/0378-1135(94)00144-l

Griesche, N., Zikos, D., Witkowski, P., Nitsche, A., Ellerbrok, H., Spiller, O. B., et al. (2008). Growth characteristic of human adenoviruses on porcine cell lines. *Virology* 373, 400–410. doi: 10.1016/j.virol.2007.12.015

Gupta, P. P. (1978). Inclusion body hepatitis in a black panther (*Panthera pardus* pardus). *Zbl Vet. Med. B.* 25, 858–860. doi: 10.1111/j.1439-0450.1978.tb01063.x

Hierholzer, J. C. (1992). Adenoviruses in the immunocompromised host. *Clin. Microbiol. Rev.* 5, 262–274. doi: 10.1128/CMR.5.3.262

Howitt, J., Bewley, M., Graziano, V., Flanangan, J., and Freimuth, P. (2003). Structural basis for variation in adenovirus affinity for the cellular receptor CAR. *J. Biol. Chem.* 278, 26208–26215. doi: 10.1074/jbc.M301492200

Jothikumar, N., Cromeans, T. L., Hill, V. R., Lu, X., Sobsey, M. D., and Erdman, D. D. (2005). Quantitative real-time PCR assays for detection of human adenoviruses and identification of serotypes 40 and 41. *Appl. Environm. Microbiol.* 71, 3131–3136. doi: 10.1128/AEM.71.6.3131-3136.2005

Kennedy, F. A. (1993). Disseminated adenovirus infection in a cat. *J. Vet. Diagn. Invest.* 5, 273–276. doi: 10.1177/104063879300500224

Kleiboeker, S. B. (1995). Sequence analysis of the fiber genomic region of a porcine adenovirus predicts a novel fiber protein. *Virus Res.* 39, 299–309. doi: 10.1016/0168-1702(95)00079-8

Kliewer, S. J., Garcia, L., Pearson, E., Soultanakis, E., Dasgupta, A., and Gaynor, R. (1989). Multiple transcriptional regulatory domains in the human immunodeficiency virus type 1 long terminal repeat are involved in basal and E1A/E1B-induced promoter activity. *J. Virol.* 63, 4616–4625.

Lakatos, B., Farkas, J., Ádám, É, Dobay, O., Jeney, C., Nász, I., et al. (2000). Serological evidence of adenovirus infection in cats. *Arch. Virol.* 145, 1029–1033. doi: 10.1007/s007050050693

Lakatos, B., Farkas, J., Ádám, É, Jarrett, O., Egberink, H. F., Bendinelli, M., et al. (1996). Data to the adenovirus infection of European cats. *Magy. Állato. Lapja Hung. Vet. J.* 51, 543–545.

Lakatos, B., Farkas, J., Egberink, H., Vennema, H., Horzinek, M. C., van Vliet, A., et al. (1997). PCR detection of adenovirus in a cat. *Magy. Állato. Lapja Hung.Vet. J.* 119, 517–519.

Lakatos, B., Farkas, J., Egberink, H., Vennema, H., Horzinek, M. C., and Benkõ, M. (1999). Detection of adenovirus hexon sequence in a cat by polymerase chain reaction. *Acta. Vet. Hung.* 47, 493–497. doi: 10.1556/AVet.47.1999.4.9

Lakatos, B., Hornyák, Á, Demeter, Z., Forgách, P., Kennedy, F., and Rusvai, M. (2017). Detection of a putative novel adenovirus by PCR amplification, sequence and phylogenetic characterization of two gene fragments from formalin-fixed paraffin-embedded tissues of a cat diagnosed with disseminated adenovirus disease. *Acta. Vet. Hung.* 65, 574–584. doi: 10.1556/004.2017.056

Lauer, K. P., Llorente, I., Blair, E., Seto, J., Krasnow, V., Purkayashtha, A., et al. (2004). Natural variation among human adenoviruses: genome sequence and annotation of human adenovirus serotype 1. *J. Gen. Virol.* 85, 2615–2625. doi: 10.1099/vir.0.80118-0

Lion, T. (2014). Adenovirus infections in immunocompetent and immunocompromised patients. *Clin. Microbiol. Rev.* 27, 441–462. doi: 10.1128/CMR.00116-13

Luiz, L. N., Gagliardi Leite, J. P., Yokosawa, J., Carneiro, B. M., Pereira Filho, E., de Mattos Oliveira, T. F., et al. (2010). Molecular characterization of adenoviruses from children presenting with acute respiratory diseases in Uberlandia. Minas Gerais, Brazil, and detection of an isolate genetically related to feline adenovirus. *Mem. Inst. Oswaldo Cruz Rio de Janeiro* 105, 712–716. doi: 10.1590/S0074-02762010000500019

Madisch, I., Hofmayer, S., Moritz, C., Grintzalis, A., Mainmueller, J., Pring-Akerblom, P., et al. (2007). Phylogenetic analysis and structural predictions of human adenovirus penton proteins as a basis for tissue-specific adenovirus vector design. *J. Virol.* 81, 8270–8281. doi: 10.1128/JVI.00048-07

Mahalingam, B., Van Agthoven, J. F., Xiong, J. P., Alonso, J. L., Adair, B. D., Rui, X., et al. (2014). Atomic basis for the species-specific inhibition of αV integrins by mAb 17E6 is revealed by the crystal structure of αVβ3 ectodomain-17E6 Fab complex. *J. Biol. Chem.* 289, 13801–13809. doi: 10.1074/jbc.M113.546929

Matthes-Martin, S., Boztug, H., and Lion, T. (2013). Diagnosis and treatment of adenovirus infection in immunocompromised patients. *Expert Rev. Anti Infect Ther.* 11, 1017–1028. doi: 10.1586/14787210.2013.836964

Maxwell, I. H., Chapman, J. T., Scherrer, L. C., Spitzer, A. L., Leptihn, S., Maxwell, F., et al. (2001). Expansion of tropism of a feline parvovirus to target a human tumor cell line by display of an alpha(v) integrin binding peptide on the capsid. *Gene Ther.* 8, 324–331. doi: 10.1038/sj.gt.3301399

Menter, D. G., Fitzgerald, L., Patton, J. T., McIntire, L. V., and Nicolson, G. L. (1995). Human melanoma integrins contribute to arrest and stabilization potential while flowing over extracellular matrix. *Immunol. Cell Biol.* 73, 575–583. doi: 10.1038/icb.1995.91

Multiple Sequence Comparison by Log-Expectation [MUSCLE] (2018). *Multiple Sequence Alignment (MUSCLE)*. Available at: http://www.ebi.ac.uk/Tools/msa/muscle/ (accessed: September 11, 2018).

Nagy, M., Nagy, É, and Tuboly, T. (2002). Sequence analysis of porcine adenovirus serotype 5 fibre gene: evidence or recombination. *Virus Genes* 24, 181–185. doi: 10.1023/A:1014580802250

Oka, M., Hitomi, T., Okada, T., Nakamura, S.-I., Nagai, H., Ohba, M., et al. (2002). Dual regulation of phospholipase D1 by protein kinase C alpha in vivo. *Biochem. Biophys. Res. Comm.* 294, 1109–1113. doi: 10.1016/S0006-291X(02)00614-9

Ongrádi, J. (1999). Identification of a feline adenovirus isolate that replicates in monkey and human cells in vitro. *Am. J. Vet. Res.* 60:1463.

Ongrádi, J., Laird, H. M., Szilágyi, J. F., Horváth, A., and Bendinelli, M. (2000). Unique morphological alterations of the HTLV-I transformed C8166 cells by infection with HIV-1. *Pathol. Oncol. Res.* 6, 27–37. doi: 10.1007/BF03032655

Ongrádi, J., Stercz, B., Kövesdi, V., Nagy, K., and Pistello, M. (2013). "Interaction of FIV with heterologous microbes in the feline AIDS model," in *Current perspectives in HIV infection*, ed. S. Saxena (Rijeka: InTech).

Orlicky, D. J., and Schaack, J. (2001). Adenovirus transduction of 3T3-L1 cells. *J. Lipid Res.* 42, 460–466.

Pedersen, N. C., Ho, E. W., Braun, M. L., and Yamamoto, J. K. (1987). Isolation of a T-lymphotropic virus from domestic cats with an immunodeficiency-like syndrome. *Science* 235, 790–793. doi: 10.1126/science.3643650

Phan, T. G., Shimizu, H., Nishimura, S., Okitsu, S., Maneekarn, N., and Ushijima, H. (2006). Human adenovirus typ1 related to feline adenovirus: evidence of interspecies transmission. *Clin. Lab.* 52, 515–518.

Pring-Åkerblom, P., Adrian, T., and Köstler, T. (1997a). PCR-based detection and typing human adenoviruses in clinical samples. *Res. Virol.* 148, 225–231. doi: 10.1016/S0923-2516(97)83992-1

Pring-Åkerblom, P., and Adrian, T. (1993). The hexon genes of adenovirus of subgenus C: comparison of the variable regions. *Res. Virol.* 144, 117–127. doi: 10.1016/S0923-2516(06)80020-8

Pring-Åkerblom, P., and Adrian, T. (1994). Type- and group-specific polymerase chain reaction for adenovirus detection. *Res. Virol.* 145, 25–35. doi: 10.1016/S0923-2516(07)80004-5

Pring-Åkerblom, P., and Adrian, T. (1995a). Characterization of adenovirus subgenus D fiber genes. *Virology* 206, 564–571. doi: 10.1016/S0042-6822(95)80073-5

Pring-Åkerblom, P., and Adrian, T. (1995b). Sequence characterization of the adenovirus 31 fibre and comparison with serotypes of subgenera A to F. *Res. Virol.* 146, 343–354. doi: 10.1016/0923-2516(96)80597-8

Pring-Åkerblom, P., and Ongrádi, J. (1980a). *Feline Adenovirus Fiber*. Available at: https://www.ncbi.nlm.nih.gov/nuccore/AY518270

Pring-Åkerblom, P., and Ongrádi, J. (1980b). *Feline Adenovirus Hexon*. Available at: https://www.ncbi.nlm.nih.gov/nuccore/AY512566.1

Pring-Åkerblom, P., Heim, A., and Trijssenaar, F. E. J. (1997b). Conserved sequences in the fibers of epidemic keratoconjunctivitis associated human adenoviruses. *Arch. Virol.* 42, 205–211. doi: 10.1007/s007050050071

Pring-Åkerblom, P., Heim, A., and Trijssenaar, F. E. J. (1998). Molecular characterization of hemagglutination domains on the fibers of subgenus D adenoviruses. *J. Virol.* 72, 2297–2304.

Pring-Åkerblom, P., Trijssenaar, F. E. J., Adrian, T., and Hoyer, H. (1999). Multiplex polymerase chain reaction for subgenus-specific detection of human adenoviruses in clinical samples. *J Med. Virol.* 58, 87–92. doi: 10.1002/(SICI)1096-9071(199905)58:1<87::AID-JMV14<3.0.CO;2-R

Pring-Åkerblom, P., Trijssenaar, F. E. J., and Adrian, T. (1995). Sequence characterization and comparison of human adenovirus subgenus B and E hexons. *Virology* 212, 232–236. doi: 10.1006/viro.1995.1474

Reddy, P. S., Nagy, É, and Derbyshire, J. B. (1995). Sequence analysis of putative pVIII, E3, and fibre regions of porcine adenovirus type 3. *Virus Res.* 36, 97–106. doi: 10.1016/0168-1702(94)00105-L

Roelvink, P. W., Lee, G. M., Einfeld, D. A., Kovesdi, I., and Wickham, T. J. (1999). Identification of a conserved receptor-binding site on the fiber proteins of CAR-recognizing adenoviridae. *Science* 286, 1568–1571. doi: 10.1126/science.286.5444.1568

Rowe, W. P., Huebner, R. J., Gilmore, L. K., Parrott, R. H., and Ward, T. G. (1953). Isolation of a cytopathogenic agent from human adenoids undergoing spontaneous degeneration in tissue culture. *Proc. Soc. Exp. Biol. Med.* 84, 570–573. doi: 10.3181/00379727-84-20714

Sarkar, D., Su, Z.-Z., Park, A.-S., Vozhilla, N., Dent, P., Curiel, D. T., et al. (2008). A cancer terminator virus eradicates both primary and distant human melanoma. *Cancer Gene Ther.* 15, 293–302. doi: 10.1038/cgt.2008.14

Sharma, P., Kolawole, A. O., Wiltshire, S. M., Frondorf, K., and Excoffon, K. J. D. A. (2012). Accessibility of the coxsackievirus and adenovirus receptor and its importance in adenovirus gene transduction efficiency. *J. Gen. Virol.* 93, 155–158. doi: 10.1099/vir.0.036269-0

Skog, J., Mei, Y. F., and Wadell, G. (2002). Human adenovirus serotypes 4p and 11p are efficiently expressed in cell lines of neural tumour origin. *J. Gen. Virol.* 83, 1299–1309. doi: 10.1099/0022-1317-83-6-1299

Stercz, B., Perlstadt, H., Nagy, K., and Ongrádi, J. (2013). Immunochemistry of adenoviruses: limitations and new horizons of gene therapy. *Acta microbial Immunol. Hung.* 60, 447–459. doi: 10.1556/AMicr.60.2013.4.6

Taharaguchi, S., Fukazawa, R., Kitazume, M., Harima, H., Taira, K., Oonaka, K., et al. (2012). Biology of fowl adenovirus type 1 infection of heterogenous cells. *Arch. Virol.* 157, 2223–2226. doi: 10.1007/s00705-012-1413-9

The HAdV Working Group (2019). Available at: http://hadvwg.emu.edu (accessed 15 March, 2019).

The International Committee on Taxonomy of Viruses [ICTV] (2003). *Taxonomy*. Available at: www.ictvdb.iacr.ac.uk (accessed April 8, 2003).

UniProt (2018). Available at: http://www.uniprot.org/taxonomy/114408 (accessed November 13, 2018).

Virus Taxonomy 2018b Release (2018). *EC50, Washington DC, July 2018*. Available at: http://talk.ictvonline.org (accessed March 15, 2019).

Wickham, T. J., Mathias, P., Cheresh, D. A., and Nemerow, G. R. (1993). Integrins alpha v beta 3 and alpha v beta 5 promote adenovirus internalization but not virus attachment. *Cell* 73, 3109–319. doi: 10.1016/0092-8674(93)90231-E

Conflict of Interest Statement: DA and LC were employed by Advanced Biotechnologies Inc. (Columbia, MD, United States).

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.

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: info@frontiersin.org | +41 21 510 17 00

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