The Geographic Structure of Viruses in the Cuatro Ciénegas Basin, a Unique Oasis in Northern Mexico, Reveals a Highly Diverse Population on a Small Geographic Scale

B. Taboada,[a] P. Isa,[a] A. L. Gutiérrez-Escolano,[b] R. M. del Ángel,[b] J. E. Ludert,[b] N. Vázquez,[d] M. A. Tapia-Palacios,[d] P. Chávez,[c] E. Garrido,[c] A. C. Espinosa,[d] L. E. Eguiarte,[e] S. López,[a] V. Souza,[e] C. F. Arias[a]

[a]Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

[b]Departamento de Infectómica y Patogénesis Molecular, Centro de Investigación y de Estudios Avanzados, Instituto Politécnico Nacional, Mexico City, Mexico

[c]Departamento de Genética y Biología Molecular, Centro de Investigación y de Estudios Avanzados, Instituto Politécnico Nacional, Mexico City, Mexico

[d]Laboratorio Nacional de Ciencias de la Sostenibilidad, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico

[e]Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico

ABSTRACT The Cuatro Ciénegas Basin (CCB) is located in the Chihuahuan desert in the Mexican state of Coahuila; it has been characterized as a site with high biological diversity despite its extreme oligotrophic conditions. It has the greatest number of endemic species in North America, containing abundant living microbialites (including stromatolites and microbial mats) and diverse microbial communities. With the hypothesis that this high biodiversity and the geographic structure should be reflected in the virome, the viral communities in 11 different locations of three drainage systems, Churince, La Becerra, and Pozas Rojas, and in the intestinal contents of 3 different fish species, were analyzed for both eukaryotic and prokaryotic RNA and DNA viruses using next-generation sequencing methods. Double-stranded DNA (dsDNA) virus families were the most abundant (72.5% of reads), followed by single-stranded DNA (ssDNA) viruses (2.9%) and ssRNA and dsRNA virus families (0.5%). Thirteen families had dsDNA genomes, five had ssDNA, three had dsRNA, and 16 had ssRNA. A highly diverse viral community was found, with an ample range of hosts and a strong geographical structure, with very even distributions and signals of endemicity in the phylogenetic trees from several different virus families. The majority of viruses found were bacteriophages but eukaryotic viruses were also frequent, and the large diversity of viruses related to algae were a surprise, since algae are not evident in the previously analyzed aquatic systems of this ecosystem. Animal viruses were also frequently found, showing the large diversity of aquatic animals in this oasis, where plants, protozoa, and archaea are rare.

IMPORTANCE In this study, we tested whether the high biodiversity and geographic structure of CCB is reflected in its virome. CCB is an extraordinarily biodiverse oasis in the Chihuahuan desert, where a previous virome study suggested that viruses had followed the marine ancestry of the marine bacteria and, as a result of their long isolation, became endemic to the site. In this study, which includes a larger sequencing coverage and water samples from other sites within the valley, we confirmed the high virus biodiversity and uniqueness as well as the strong biogeographical diversification of the CCB. In addition, we also analyzed fish intestinal contents, finding that each fish species eats different prey and, as a result, presents different viral compositions even if they coexist in the same pond. These facts highlight the high and novel virus diversity of CCB and its “lost world” status.

Received 26 February 2018 **Accepted** 21 March 2018

Accepted manuscript posted online 6 April 2018

Citation Taboada B, Isa P, Gutiérrez-Escolano AL, del Ángel RM, Ludert JE, Vázquez N, Tapia-Palacios MA, Chávez P, Garrido E, Espinosa AC, Eguiarte LE, López S, Souza V, Arias CF. 2018. The geographic structure of viruses in the Cuatro Ciénegas Basin, a unique oasis in northern Mexico, reveals a highly diverse population on a small geographic scale. Appl Environ Microbiol 84:e00465-18. https://doi.org/10.1128/AEM.00465-18.

Editor Harold L. Drake, University of Bayreuth

Address correspondence to V. Souza, souza.valeria2@gmail.com, or C. F. Arias, arias@ibt.unam.mx.

KEYWORDS bacteriophages, shotgun metagenomics, viruses

Viruses infect all known forms of life and are the most abundant and diverse biological entities of all Earth's ecosystems, including extreme environments such as the oceanic basement (1). Viruses play an integral role in the life cycle of their hosts, affecting not only their population demography but also their genetic diversity, thus strongly influencing the function, ecology, and evolution of complete communities and ecosystems (2). For instance, the ocean cyanophage virus-host dynamics affect different critical aspects of the ocean ecology, from photosynthesis to the geochemical cycle of phosphate and nitrogen (3, 4).

Massive next-generation sequencing (NGS) strategies and metagenomic analyses have been successfully used to overcome some of the limitations of classical methods for virus detection and characterization in diverse ecosystems. Metagenomic studies of the oceans (5–14) and other environments (15–19) have shown that viruses indeed are diverse and abundant in most habitats. Viruses have been found even below the basaltic basement (1), where most of them were related to *Archaea* and were hard to classify. This reflects a constant technical caveat in all viromic studies, since databases are incomplete and a large number of reads remains uncharacterized and are referred to as "dark matter," correlating with the "dark biodiversity" of their hosts (20). This lack of reference sequences in databases is even more pronounced in studies that include RNA viruses (7, 15).

The Cuatro Ciénegas Basin (CCB) is located in the Chihuahuan desert in the Mexican state of Coahuila; it has been characterized as a site with high biological diversity despite its extreme oligotrophic conditions. It harbors the greatest number of endemic species of any place in North America, including abundant living microbialites (including stromatolites), microbial mats, and diverse microbial communities in all of the explored biomes (21). The area is a small setting of less than 840 km^2 surrounded by mountains rising up to 3,500 m above mean sea level, and it is divided into seven major and permanent drainage systems that form hundreds of ponds and small lakes (locally called "pozas") with crystalline turquoise blue water (22).

Previous studies have described the great biodiversity of vertebrates, invertebrates, and plants (22–24), as well as of prokaryotes (25–29), found in CCB. This high biodiversity is particularly interesting since CCB is very poor in nutrients (30, 31) and has a low primary productivity (31). The diversity of CCB has been associated mainly with local adaptations and has shown a high divergence between each of its sequenced microbialite communities and the rest of the world. Also, marine ancestry was described using either environmental 16S rRNA gene clones (25) or cultured bacteria (27) and later confirmed by genome sequencing (32, 33) and metagenomics (34). These results have led authors to refer to CCB as a "lost world," where ancient marine lineages, lost to the rest of the planet, still persist (29).

Little is known about the viral diversity in CCB, even though it represents a hot spot of endemicity for all taxon levels. A single previous study that characterized the phage content in modern stromatolites and thrombolites of two of the hydrological systems of CCB, Pozas Azules and Rio Mezquites, reported an extraordinarily high viral diversity associated with the Pozas Azules stromatolite (35). In addition, genetic similarities between the Gulf of Mexico, the Sargasso Sea, and the CCB phage communities were found, despite the fact that these environments have not been in contact for the last 35 million years, when in the late Eocene the rise of the Mexican Sierra Madre Oriental isolated CCB from the Gulf of Mexico.

In this study, we used high-coverage NGS methods to analyze the viral communities of three different aquatic systems in CCB not previously studied, looking for both eukaryotic and prokaryotic RNA and DNA viruses. The first site, Churince, has been heavily impacted by the overexploitation of the deep aquifer (36). The second site, La Becerra, was used as a recreational pond for decades but is now under recovery, since it was closed to the public in 2009. The third site, Pozas Rojas, is a naturally fluctuating

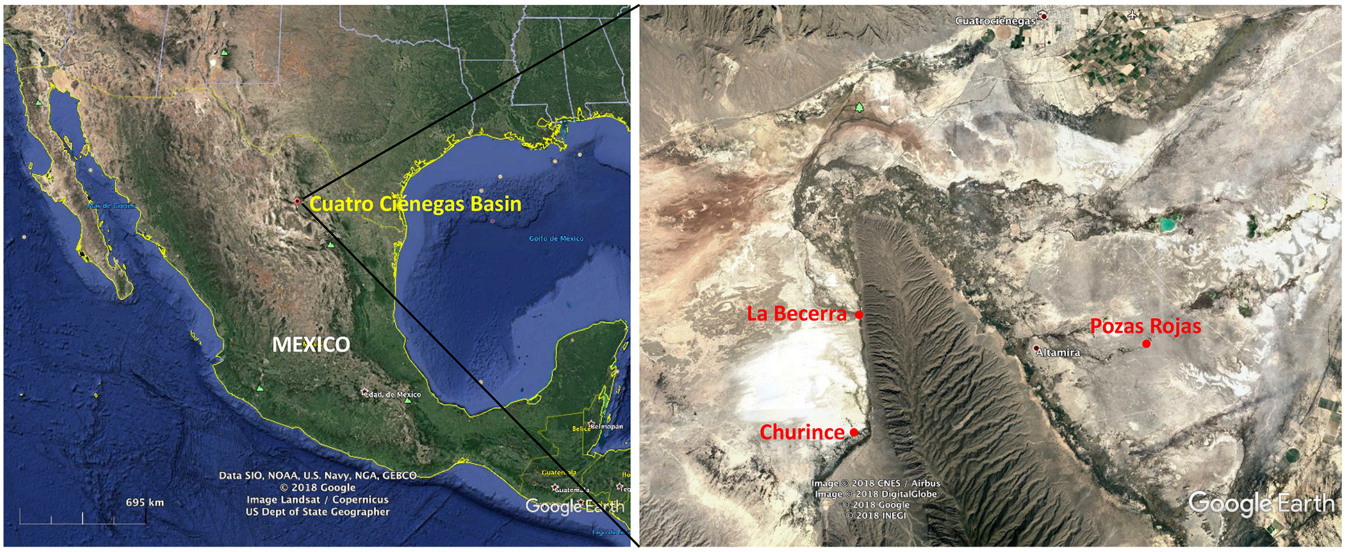


FIG 1 Map of the study area, Cuatro Ciénegas Basin in the state of Coahuila, Mexico.

environment with no human disturbance, but it has gone through a succession process after Hurricane Alex moved a large amount of water to these sinkholes in the spring of 2010. We also characterized the viruses present in the intestinal contents of three different fish species of Churince.

The metagenomic analysis showed a highly diverse viral community with an ample range of hosts and a strong geographical structure, with distributions that in general are very even. In addition, phylogenetic analysis based on different viral proteins indicated strong signals of endemicity for several different virus families. Finally, as expected, due to the high bacterial diversity, the majority of viruses found were bacteriophages; a high diversity and abundance of viruses related to algae was also present, an unexpected finding since algae are not visually evident in the ponds analyzed. Animal viruses were also frequently found, in correspondence to the large diversity of aquatic animals in this oasis, where plants, protozoa, and archaea are rare.

RESULTS

Overview of the virome data sets. To study the viral communities of CCB, water samples were collected from 11 different locations of three drainage systems: Churince, La Becerra, and Pozas Rojas (Fig. 1). Fish intestinal contents (FIC) from three different species that coexist in the Churince spring were also obtained. The total number of sequence reads obtained for each sample (between 4 and 13 million) is shown in Table 1; the average length of reads was 125 nucleotides. After quality trimming, size filtering, and collapsing duplicate reads, 83.7% and 73% of the water and FIC sequence reads remained, respectively, and were further analyzed.

Viral taxonomic composition of water samples. The taxonomic composition was determined through comparison of valid unique sequence reads against sequences of bacterial, fungal, and viral origin in the GenBank nucleotide database and all-nonredundant-protein (nr) database using BLASTN and BLASTX, respectively (see Materials and Methods). Most sequence reads (64.6% to 83.2%) obtained from the water samples showed no significant similarity to any known sequence deposited in the GenBank database (see Fig. S1 in the supplemental material), suggesting that a large proportion of the microorganisms in the environment is unknown and that these no-hit reads could be derived from either uncharacterized viruses or other taxa. Out of all the reads, an average of 3.8% (1.6% to 10%) was identified as viral by the combined nucleotide and protein analysis of MEGAN 5, representing 144 genera (Table S1) and 1,434 different known viral species (Table S2) that belong to 37 different viral families (Table 2).

TABLE 1 Samples used in the study and overview of the total number of sequence reads and valid reads obtained through NGS

Sample site	Collection date (day/mo/yr)	Vol collected	Final concn (ml)	Sample description	Total no. of reads	No. (%) of valid reads
L2	06/02/14	9.8 liters	77	Water from Churince	7,962,496	7,505,587 (94.3)
L4	06/03/14	8.0 liters	47	Water from Churince	9,974,898	8,251,324 (82.7)
L5	07/03/14	11.4 liters	70	Water from Churince	8,269,314	7,456,526 (90.2)
L9	07/03/14	11.4 liters	72	Water from Churince	4,247,137	3,853,484 (90.7)
L10	08/03/14	4.3 liters	72	Water from Churince	7,017,080	6,137,055 (87.5)
BE	08/03/14	10.8 liters	55	Water from La Becerra	8,183,308	6,973,697 (85.2)
PR1	09/03/14	9.5 liters	52	Water from Pozas Rojas	4,805,076	4,353,067 (90.6)
PR3	09/03/14	9.8 liters	65	Water from Pozas Rojas	5,203,288	4,247,703 (81.6)
PR4	09/03/14	7.0 liters	75	Water from Pozas Rojas	13,320,976	1,035,9155 (77.8)
PR7	09/03/14	9.5 liters	57	Water from Pozas Rojas	7,266,635	6,133,145 (84.4)
PR9	09/03/14	9.6 liters	57	Water from Pozas Rojas	5,908,946	5,461,242 (92.4)
HG[a]	30/06/14	2 ml		FIC of *Hemichromis guttatus*	7,531,002	5,870,049 (77.9)
GM[a]	30/06/14	1.5 ml		FIC of *Gambusia marshi*	7,311,105	4,957,444 (67.8)
CB[a]	30/06/14	0.7 ml		FIC of *Cyprinodon bifasciatus*	8,172,051	5,982,252 (73.2)

[a] Collected in Churince. *H. guttatus* (HG) is an invasive fish, and *G. marshi* (GM) and *C. bifasciatus* (CB) are fish endemic to CCB.

Analysis of the identified virus families, based on their type of genome, revealed that the most frequent were families with double-stranded DNA (dsDNA) and single-stranded RNA (ssRNA) genomes. Thirteen families had dsDNA genomes, 5 had ssDNA (1 of negative-sense polarity, 3 of positive-sense polarity, and 1 ambisense), 3 were families with dsRNA genomes, and 16 had ssRNA (1 of negative-sense polarity and the rest with positive-sense genomes) (Fig. 2A). However, based on the number of sequence reads, dsDNA virus families were by far the most abundant (72.5% of reads) (Fig. 2B), followed by viruses with ssDNA, which represented 2.9% of the reads, and ssRNA and dsRNA virus families, representing only 0.5% of the reads each. Interestingly, an average of 22% of the reads in each sample could not be assigned to any previously described viral family (Fig. 2B), and these nonassigned sequences were similar to unclassified viruses and/or to viruses found in diverse environmental samples.

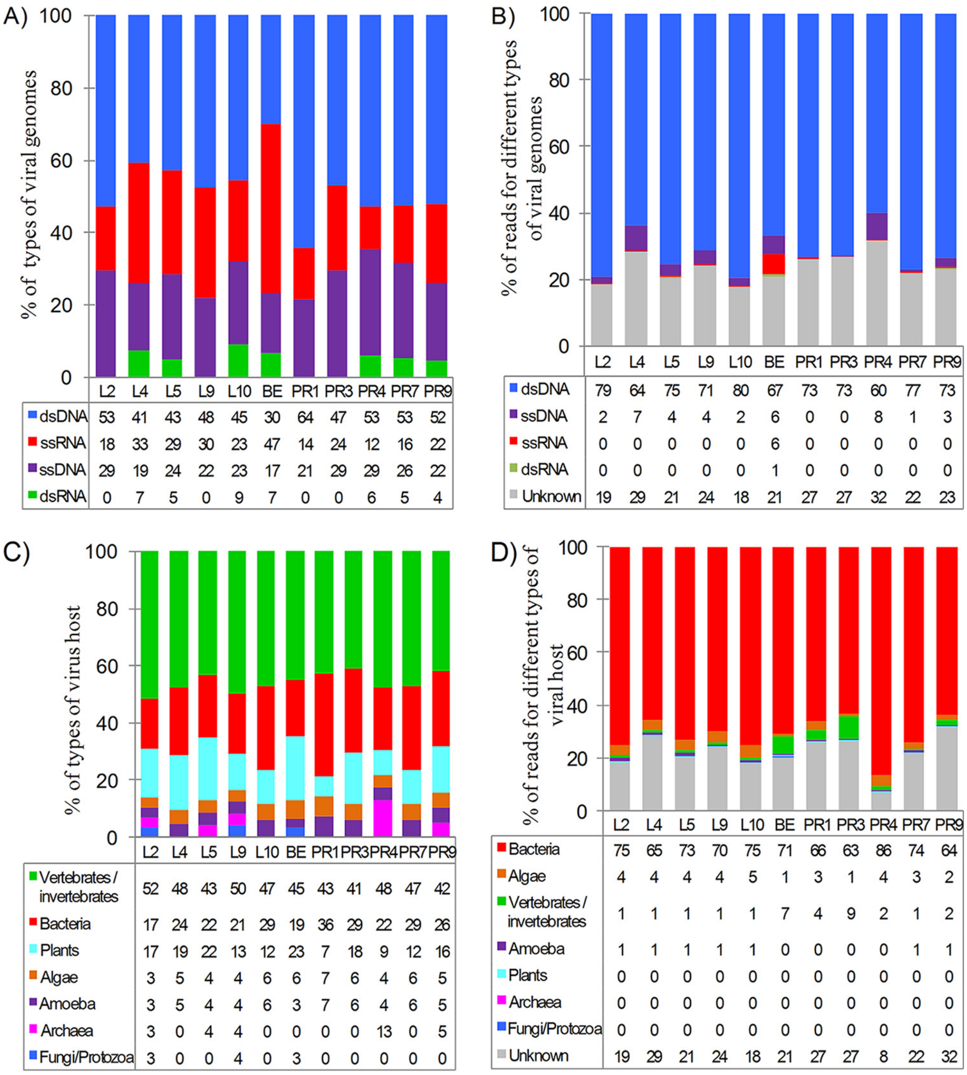
The viral families identified in this work have been reported to have a wide range of hosts. About half of them infect vertebrates or invertebrates, followed by those that infect bacteria (23.8%), plants (15.7%), algae (5.1%), amoebae (4.7%), archaea (2.9%), and fungi/protozoa (1.4%) (Fig. 2C). In accordance with the observation that most of the sequence reads were found to belong to dsDNA viruses, particularly bacteriophages, 71.6% of the reads were related to bacterial hosts (Fig. 2D). Phages from 6 virus families were found in these samples, with *Myoviridae*, *Siphoviridae*, and *Podoviridae* being the most abundant; one ssDNA family, *Microviridae*, was also found, although with a lower abundance (Fig. 3A and Table 2). Other, less abundant bacteriophage families were the dsDNA *Tectiviridae* (on average, 0.05% of all reads), present in all samples, and the positive-sense ssRNA [(+)ssRNA] *Leviviridae* family (0.003%), found only in La Becerra.

The next dominant reads present in all samples, at the family level (Table 2), share identity to eukaryotic dsDNA viruses from the *Phycodnaviridae* family, which infects algae, and from the *Iridoviridae* family, which infects invertebrates and vertebrates. Viruses from the *Mimiviridae* family, which infects amoebae, were also found in all samples. The amino acid identities to known viral proteins in all virus sequences were less than 60%, suggesting that most of these viruses are novel members of the identified virus families, in agreement with the uniqueness of CCB. Besides these major viral families, the remaining 28 families made up, on average, only 1.5% of the total reads (Fig. 3A and Table 2). Overall, we identified (i) 6 virus families that infect plants, (ii) 4 families that infect invertebrates, including aphids, leafhoppers, flies, bees, ants, silkworms, wasps, moths, butterflies, crustacea, and arthropods, (iii) 8 families that infect vertebrates (the oasis is diverse in fishes, reptiles, birds, and amphibians), (iv) 5 families that infect both invertebrates and vertebrates, (v) 3 virus families that infect archaea, (vi) 1 family that infects algae, and finally, (vii) 1 dsDNA family that infects

TABLE 2 Distribution of viral reads at family level from water and fish samples[a]

		No. of reads by sample source														
		Churince					La Becerra	Pozas Rojas					FIC			
Virus family	Genome type	L2	L4	L5	L9	L10	(BE)	PR1	PR3	PR4	PR7	PR9	HG	GM	CB	Host
Alphatetraviridae	(+)ssRNA	0	0	0	0	0	15	0	0	0	0	0	0	0	0	Invertebrates
Ampullaviridae	dsDNA	0	0	0	3	0	0	0	0	3	0	0	0	0	0	Archaea
Anelloviridae	dsDNA	13	0	18	0	0	0	0	0	0	0	0	0	0	0	Vertebrates
Asfarviridae	dsDNA	58	33	16	20	24	8	5	0	13	7	21	6	0	0	Vertebrates
Astroviridae	(+)ssRNA	31	0	41	5	0	3	3	5	0	0	0	23	29,286	15	Vertebrates
Baculoviridae	dsDNA	11	10	11	4	18	5	3	3	30	7	24	11	27	2	Invertebrates
Bunyaviridae	(−)ssRNA	0	0	0	0	3	0	0	0	0	0	77	12	0	4	Vertebrates
Caliciviridae	(+)ssRNA	0	0	0	0	0	5	0	0	7	0	0	100	9	51	Vertebrates
Circoviridae	(−)ssDNA	607	366	403	263	212	185	161	38	2,113	916	126	1,492	71	193	Vertebrates
Closteroviridae	(+)ssRNA	0	3	4	0	0	6	0	0	0	0	0	0	0	0	Plants
Dicistroviridae	(+)ssRNA	238	346	55	46	22	4,228	55	90	1,377	365	28	124,264	163,059	579,409	Invertebrates
Fuselloviridae	dsDNA	3	0	0	0	0	0	0	0	7	0	7	0	0	0	Archaea
Geminiviridae	(+)ssDNA	45	15	26	6	4	16	0	5	33	10	16	0	0	16	Plant
Hepeviridae	(+)ssRNA	47	0	0	0	0	0	0	0	0	0	0	48	0	334	Vertebrates
Iflaviridae	(+)ssRNA	15	3	0	3	3	53	0	3	14	0	0	663	11,988	1,725	Invertebrates
Iridoviridae	dsDNA	442	547	388	473	731	276	7,695	25,587	917	1,521	595	110	527	136	Vertebrates/invertebrates
Leviviridae	(+)ssRNA	0	0	0	0	0	52	0	0	0	0	0	0	0	0	Bacteria
Luteoviridae	(+)ssRNA	0	0	0	0	0	25	0	0	0	0	0	746	43	14	Plants
Marnaviridae	(+)ssRNA	0	0	0	0	0	17	0	0	0	0	0	4	0	16	Algae
Microviridae	(+)ssDNA	10,148	7,352	4,710	2,115	2,822	4,918	182	551	8,153	11,404	804	75,319	4,353	9,266	Bacteria
Mimiviridae	dsDNA	1,311	2,844	790	1,465	2,296	356	579	147	1,090	811	1,317	260	978	419	Amoeba
Myoviridae	dsDNA	29,914	66,231	27,438	30,402	48,351	36,933	47,503	68,508	142,878	38,315	39,874	190,649	70,571	5,703	Bacteria
Nanoviridae	(+)ssDNA	108	47	58	28	25	12	60	14	122	87	11	25	9	20	Plant
Nodaviridae	(+)ssRNA	13	8	10	12	0	43	0	0	0	0	0	682	0	0	Vertebrates/invertebrates
Partitiviridae	dsRNA	0	0	0	0	0	4	0	0	0	0	0	693	78	120	Fungi/plants
Parvoviridae	(±)ssDNA	59	54	16	6	4	29	0	5	8	4	5	1,979	8,589	5,311	Vertebrates/invertebrates
Permutotetraviridae	(+)ssRNA	0	0	0	0	0	0	0	0	0	0	0	84	548	0	Invertebrates
Phycodnaviridae	dsDNA	5,789	8,932	5,493	5,346	6,270	682	6,254	3,801	12,894	2,950	4,537	674	451	719	Algae
Picobirnaviridae	dsRNA	20	4	0	3	0	691	0	0	69	19	2	33	0	80	Vertebrates
Picornaviridae	(+)ssRNA	27	0	7	3	0	145	0	0	83	0	10	1,082	1,321	5,227	Vertebrates/invertebrates /plants
Podoviridae	dsDNA	16,249	16,998	15,681	9,873	17,962	8,706	37,191	82,860	17,284	14,849	21,339	51,960	129,271	4,715	Bacteria
Reoviridae	dsRNA	4	0	0	10	0	0	0	0	0	0	0	45	6	15	Vertebrates/invertebrates /plants/fungi
Secoviridae	(+)ssRNA	7	17	0	0	0	110	0	0	189	34	0	3,815	191	29,043	Vertebrates/invertebrates
Siphoviridae	dsDNA	41,219	66,330	40,318	44,066	53,920	14,367	51,441	36,394	99,084	37,065	67,332	40,288	87,701	6,080	Bacteria
Tectiviridae	dsDNA	195	142	141	83	186	113	3	3	41	42	9	6	6	0	Bacteria
Tombusviridae	(+)ssRNA	44	21	24	0	0	256	0	7	0	0	24	143	3,412	5,871	Plant
Totiviridae	dsRNA	0	0	0	0	0	0	0	0	0	0	0	72	86	16	Protozoa/fungi
Turriviridae	dsDNA	0	0	4	0	0	0	0	0	7	0	0	0	0	0	Archaea
Tymoviridae	(+)ssRNA	0	0	0	0	0	0	0	0	0	0	0	0	1,796	0	Plants
Virgaviridae	(+)ssRNA	20	0	66	66	0	114	0	0	0	0	0	0	8	39	Plants

[a] The read count is the result of the combined stringency analysis of BLASTx and BLASTn MEGAN taxa assignment collapsed at this level. HG, *Hemicromis gutatus* fish; GM, *Gambusia marshi* fish; CB, *Cyprinon bifacetus* fish; FIC, fish intestinal contents. A total of 40 virus families are listed, 37 from water samples and 33 from FIC.



A)	L2	L4	L5	L9	L10	BE	PR1	PR3	PR4	PR7	PR9
dsDNA	53	41	43	48	45	30	64	47	53	53	52
ssRNA	18	33	29	30	23	47	14	24	12	16	22
ssDNA	29	19	24	22	23	17	21	29	29	26	22
dsRNA	0	7	5	0	9	7	0	0	6	5	4

B)	L2	L4	L5	L9	L10	BE	PR1	PR3	PR4	PR7	PR9
dsDNA	79	64	75	71	80	67	73	73	60	77	73
ssDNA	2	7	4	4	2	6	0	0	8	1	3
ssRNA	0	0	0	0	0	6	0	0	0	0	0
dsRNA	0	0	0	0	0	1	0	0	0	0	0
Unknown	19	29	21	24	18	21	27	27	32	22	23

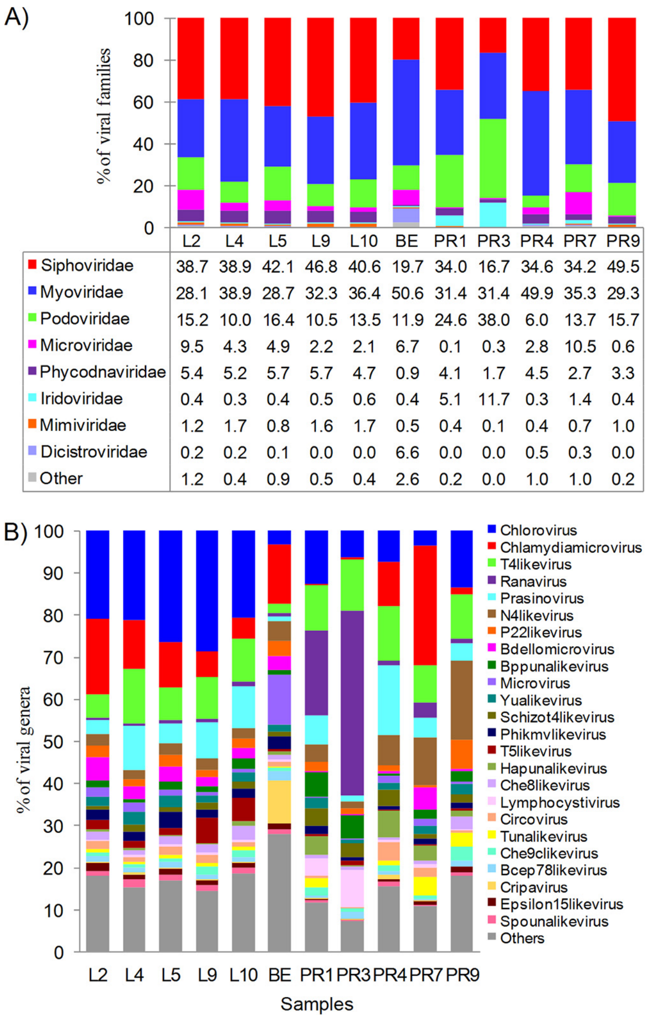
C)	L2	L4	L5	L9	L10	BE	PR1	PR3	PR4	PR7	PR9
Vertebrates / invertebrates	52	48	43	50	47	45	43	41	48	47	42
Bacteria	17	24	22	21	29	19	36	29	22	29	26
Plants	17	19	22	13	12	23	7	18	9	12	16
Algae	3	5	4	4	6	6	7	6	4	6	5
Amoeba	3	5	4	4	6	3	7	6	4	6	5
Archaea	3	0	4	4	0	0	0	0	13	0	5
Fungi/Protozoa	3	0	0	4	0	3	0	0	0	0	0

D)	L2	L4	L5	L9	L10	BE	PR1	PR3	PR4	PR7	PR9
Bacteria	75	65	73	70	75	71	66	63	86	74	64
Algae	4	4	4	4	5	1	3	1	4	3	2
Vertebrates / invertebrates	1	1	1	1	1	7	4	9	2	1	2
Amoeba	1	1	1	1	1	0	0	0	0	1	1
Plants	0	0	0	0	0	0	0	0	0	0	0
Archaea	0	0	0	0	0	0	0	0	0	0	0
Fungi/Protozoa	0	0	0	0	0	0	0	0	0	0	0
Unknown	19	29	21	24	18	21	27	27	8	22	32

FIG 2 Relative abundance of viral genome types and virus hosts identified in water samples of CCB based on genome types of the viruses identified (A), reads assigned to the type of virus genome (B), viruses based on the type of host they infect (C), and reads based on the host organism of the identified viruses (D). The sequences assigned to unclassified viruses or viruses found in environmental samples for which the host organism has not been described are placed under "Unknown." L2 to L10 are samples from Churince, BE samples are from La Becerra, and PR3 to PR9 samples are from Pozas Rojas.

amoebae. The viral diversity found is indeed a reflection of the ecosystem biodiversity of virus hosts in the CCB.

Virus genera and species in water samples. At the genus level, the most abundant genera found across all samples were *Chlorovirus* and *Prasinovirus*, which infect algae, *Chlamydiamicrovirus*, the T4-like virus, and the N4-like virus, which infect bacteria, and the *Ranavirus*, which infect invertebrates (Fig. 3B and Table S1). Regarding the most abundant viral species (Table S2), it was interesting to find that La Becerra had no individual virus noticeably overrepresented, while the more fluctuating systems, Churince and Pozas Rojas, had dominant but different virus species. In Pozas Rojas, the most abundant species were *Methylophilales* phage HIM624-A, *Colwellia* phage 9A, *Synechococcus* phage S-PM2, *Pelagibacter* phage HTVC008M, and *Acinetobacter* phage phiAC-1, while in Churince the most predominant species were *Cellulophaga* phage phiST, *Methylophilales* phage HIM624-A, uncultured Mediterranean phage uvMED, and *Flavobacterium* phage 11b (Table S1). Interestingly, all of the phages mentioned above have been isolated from seawater, suggesting that viruses related to marine lineages are still abundant in the two CCB systems studied.

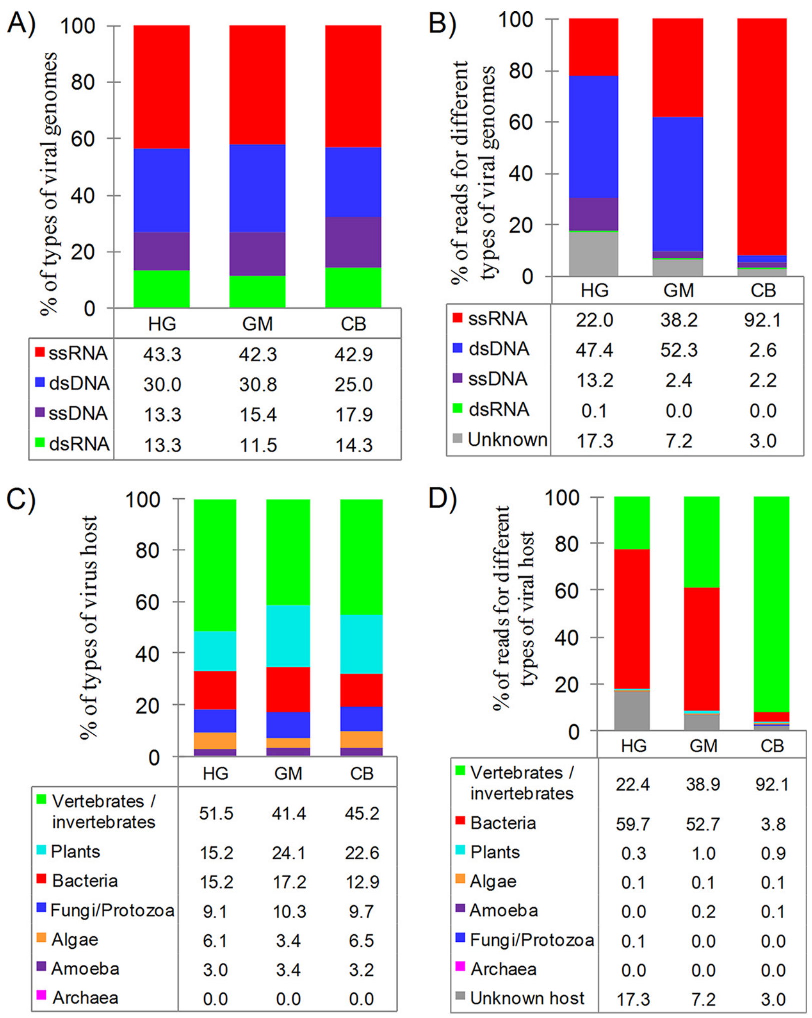


	L2	L4	L5	L9	L10	BE	PR1	PR3	PR4	PR7	PR9
🟥 Siphoviridae	38.7	38.9	42.1	46.8	40.6	19.7	34.0	16.7	34.6	34.2	49.5
🟦 Myoviridae	28.1	38.9	28.7	32.3	36.4	50.6	31.4	31.4	49.9	35.3	29.3
🟩 Podoviridae	15.2	10.0	16.4	10.5	13.5	11.9	24.6	38.0	6.0	13.7	15.7
Microviridae	9.5	4.3	4.9	2.2	2.1	6.7	0.1	0.3	2.8	10.5	0.6
🟪 Phycodnaviridae	5.4	5.2	5.7	5.7	4.7	0.9	4.1	1.7	4.5	2.7	3.3
Iridoviridae	0.4	0.3	0.4	0.5	0.6	0.4	5.1	11.7	0.3	1.4	0.4
🟧 Mimiviridae	1.2	1.7	0.8	1.6	1.7	0.5	0.4	0.1	0.4	0.7	1.0
Dicistroviridae	0.2	0.2	0.1	0.0	0.0	6.6	0.0	0.0	0.5	0.3	0.0
Other	1.2	0.4	0.9	0.5	0.4	2.6	0.2	0.0	1.0	1.0	0.2

FIG 3 Relative abundance of reads assigned to viral taxa in water samples. (A) Analysis of viral families. (B) Analysis of viral genera. Viral families and genera, whose sum of relative abundance across all samples was less than 10% and 5%, respectively, are presented as "Others." L2 to L10 are samples from Churince, BE samples are from La Becerra, and PR3 to PR9 samples are from Pozas Rojas.

Viral taxonomic composition in fish intestinal contents. The viruses present in the intestinal content of fish species *Hemichromis guttatus*, *Gambusia marshi*, and *Cyprinodon bifasciatus* (CB) were also analyzed. *H. guttatus* is an invasive species from Africa, known as jewelfish, which was probably released in CCB from an aquarium; the other 2 species are endemic to the basin. We detected a total of 1,264 different known viral species, belonging to 151 genera, and 33 viral families in the FIC (Table 2 and Tables S1 and S2) that were comprised of 9 dsDNA, 5 ssDNA, 4 dsRNA, and 15 ssRNA viral genome types (Fig. 4A). At variance with the water samples, the reads from the FIC were assigned more equally among all types of viral genomes, with ssRNA families being the most abundant (52.8%), followed by dsDNA (32.3%), ssDNA (5.8%), and dsRNA (0.2%) families (Fig. 4B). Interestingly, even within this selective environment, 8.9% of reads were found to share identity with unclassified viruses and/or unidentified viruses found in diverse environmental samples.

The viruses identified in FIC were found to have a wide range of hosts (Fig. 4C), including, in order of frequency, vertebrates/invertebrates, plants, bacteria, fungi/protozoa, algae, and amoebae. In contrast to the water samples (Fig. 4D), on average 36.9% of the sequences were assigned to bacteriophage families, while almost half of the sequences were assigned to invertebrate virus families, a reflection of the fish diet.



A)

	HG	GM	CB
🟥 ssRNA	43.3	42.3	42.9
🟦 dsDNA	30.0	30.8	25.0
🟪 ssDNA	13.3	15.4	17.9
🟩 dsRNA	13.3	11.5	14.3

B)

	HG	GM	CB
🟥 ssRNA	22.0	38.2	92.1
🟦 dsDNA	47.4	52.3	2.6
🟪 ssDNA	13.2	2.4	2.2
🟩 dsRNA	0.1	0.0	0.0
⬜ Unknown	17.3	7.2	3.0

C)

	HG	GM	CB
🟩 Vertebrates / invertebrates	51.5	41.4	45.2
🟦 Plants	15.2	24.1	22.6
🟥 Bacteria	15.2	17.2	12.9
🟦 Fungi/Protozoa	9.1	10.3	9.7
🟧 Algae	6.1	3.4	6.5
🟪 Amoeba	3.0	3.4	3.2
🟪 Archaea	0.0	0.0	0.0

D)

	HG	GM	CB
🟩 Vertebrates / invertebrates	22.4	38.9	92.1
🟥 Bacteria	59.7	52.7	3.8
🟦 Plants	0.3	1.0	0.9
🟧 Algae	0.1	0.1	0.1
🟪 Amoeba	0.0	0.2	0.1
🟦 Fungi/Protozoa	0.1	0.0	0.0
🟪 Archaea	0.0	0.0	0.0
⬜ Unknown host	17.3	7.2	3.0

FIG 4 Relative abundance of viral genome types and virus hosts identified in fish intestinal content. (A) Genome types of the viruses identified. (B) Reads assigned to the type of virus genome. (C) Viruses based on the type of host they infect. (D) Reads based on the host organism of the identified viruses. The sequences assigned to unclassified viruses or viruses found in diverse environmental samples, for which the host organism has not been described, are placed under "Unknown." HG, *Hemicromis gutatus*; GM, *Gambusia marshi*; CB, *Cyprinodon bifasciatus.*

For example, sample *C. bifasciatus* had 91.2% of the reads assigned to viruses in the positive-sense ssRNA *Dicistroviridae* family, which infects insects (Table 2 and Fig. 5A). This virus family was also the most abundant in *G. marshi* (29.4%) and second most abundant in *H. guttatus* (20.7%). The next most abundant reads were sequences with identity to dsDNA phages (3.8% to 59.7%), corresponding to the families *Myoviridae*, *Podoviridae*, *Siphoviridae*, and *Tectiviridae*, and ssDNA phages of the *Microviridae* family (Fig. 5A). Other less abundant viral families were also detected (Table 2). Finally, and in contrast to the water samples, no archaeal virus families were found in the FIC samples, suggesting that the sampled fish species do not feed on anoxic sediments.

Viral genera and species in fish intestinal contents. At the genus level (Fig. 5B), consistent with the fish diet, the most relatively abundant viral genera across all FIC samples were viruses that share identity to *Cripavirus*, *Aparavirus*, and *Iflavirus*, whose hosts are principally insects (Fig. 4D); *Microvirus*, *Chlamydiamicrovirus*, *T5*-like virus, and *Phieco32*-like virus, whose hosts are bacteria; and *Waikavirus*, whose natural hosts are plants. In this sense, two of the most abundant viral species (Table S2) belonged to the *Dicistroviridae* family: cricket paralysis virus and aphid lethal paralysis virus, whose hosts are insects. Also, three phages were commonly found, *Yersinia* phage PY100, *Bacillus*

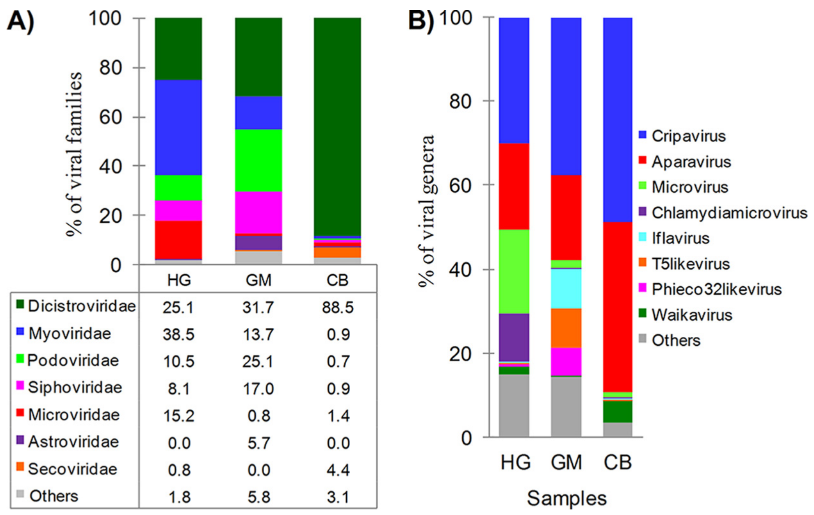


FIG 5 Relative abundance of reads assigned to viral taxa in fish intestinal content. (A) Viral families. (B) Viral genera. Viral families and species whose sum of relative abundance across all samples was less than 5% and 10%, respectively, are presented as “Others.” HG, *Hemicromis gutatus*; GM, *Gambusia marshi*; CB, *Cyprinodon bifasciatus*.

phage BCD7, and *Salmonella* phage 7-11; the host bacteria for these genera are in many cases pathogenic for vertebrates, in particular for fishes and reptiles, which are abundant in the ponds.

Functional community composition. The SEED and InterPro2GO functional profile of all samples was explored using MEGAN 5. Using the SEED database, only 12.4% to 24.5% of the reads from water and 0.8% to 15.9% from fish content samples could be functionally classified (Table 3). The dominant annotation of SEED belonged to the subsystem “Phages, prophages, transposable elements, plasmids” (46.5 to 56.5%). “Phage capsid proteins” and “Phage head and packaging” were the largest part of this group, while “Phage family *Inoviridae*,” “T4-like cyanophage core proteins,” and “T7-like

TABLE 3 Distribution of viral reads into top predicted functional SEED and InterPro2GO categories[a]

	Result by sample source													
Parameter	L2	L4	L5	L9	L10	BE	PB1	PB3	PB4	PB7	PB9	HG	CM	CB
SEED metabolic analysis														
No. of reads assigned	24,752	52,705	19,544	23,616	33,320	20,365	25,135	38,766	34,403	22,860	16,219	119,279	90,821	6,611
% assigned	20.5	29.7	19.8	25.9	26.9	14.3	13.6	13.4	10.0	17.1	8.97	23.4	17.9	1.0
Relative abundance (%) of SEED subsystems														
Cofactors, vitamins, prosthetic group, pigments	9.5	16.3	12.8	15.9	13.0	1.2	11.5	1.0	7.2	2.6	6.4	0.1	5.5	0.9
DNA metabolism	1.1	0.4	1.1	0.5	0.9	1.5	2.4	3.2	2.8	1.9	2.5	2.2	1.1	0.7
Nucleosides and nucleotides	4.3	6.0	6.1	6.4	7.8	25.9	6.6	5.6	8.8	3.0	12.1	0.3	10.6	4.2
Phages, prophages, transposable elements, plasmids	66.6	46.5	55.8	48.4	51.0	58.8	65.8	84.5	71.2	87.5	69.8	95.5	72.6	89.5
Regulation and cell signaling	5.4	11.1	7.4	10.2	9.5	0.1	1.6	0.4	1.2	0.6	1.0	0.0	1.3	0.1
RNA metabolism	8.4	15.3	11.6	15.0	12.1	0.2	3.0	0.4	1.9	0.9	1.8	0.0	0.1	0.3
Others	5.1	4.7	5.5	3.7	6.0	15.3	9.2	4.9	7.1	3.7	6.8	3.3	13.6	5.0
InterPro2GO functional analysis														
No. of reads assigned	60,623	85,927	41,772	38,179	49,409	69,669	75,760	176,572	127,094	52,038	44,648	329,081	166,713	467,215
% assigned	50.4	48.5	42.4	42.0	39.9	49.0	41.2	61.2	37.1	39.1	24.7	64.7	33.0	72.1
Relative abundance (%) of GO subsystems														
Nucleotide binding (0000166)	4.5	3.8	5.9	5.0	5.0	5.9	7.1	14.0	10.0	6.8	7.1	12.5	15.5	23.2
DNA metabolic process (0006259)	5.3	4.5	7.9	6.7	6.3	3.2	10.6	18.0	6.3	4.6	6.6	3.8	6.0	0.1
Biosynthetic process (0009058)	3.5	5.3	4.3	5.1	5.3	5.3	11.0	15.2	4.1	4.6	6.3	3.5	6.2	0.7
Transferase activity (0016740)	6.5	6.3	8.2	7.5	8.0	9.4	12.6	20.5	7.7	7.3	11.1	14.8	23.1	23.9
Hydrolase activity (0016787)	3.1	4.8	3.8	4.6	4.7	3.6	7.1	10.2	2.2	7.8	4.2	13.1	15.9	23.9
Others	77.1	75.3	69.9	71.1	70.7	72.1	51.6	22.1	69.6	68.9	64.7	52.3	33.3	28.2

[a] L2, L4, L5, L9, and L10 correspond to Churince water samples, BE to La Becerra, and PR1, PR3, PR4, PR7, and PR9 to Pozas Rojas. HG, *Hemicromis gutatus* fish; GM, *Gambusia marshi* fish; CB, *Cyprinodon bifasciatus* fish. Level 1 and 2 subsystems were used for SEED and InterPro2GO, respectively.

TABLE 4 Diversity and evenness of virus species in water and FIC samples determined by PHACCS best rank abundance model prediction

Sample or region	Richness (no. of genotypes)	Shannon index	Evenness	Model	Error
Sample					
L2	34,678	9.1	0.86	Power	274
L4	11,795	7.6	0.81	Lognormal	334
L5	27,904	8.6	0.854	Power	309
L9	20,000	8.2	0.83	Lognormal	438
L10	15,302	7.4	0.77	Lognormal	286
BE	29,401	9.2	0.89	Power	192
PR1	2,765	6.9	0.88	Lognormal	500
PR3	929	6.1	0.89	Lognormal	366
PR4	2,235	6.7	0.86	Lognormal	1,443
PR7	3,663	6.6	0.82	Lognormal	484
PR9	3,314	6.5	0.81	Lognormal	416
HG	642	5.9	0.91	Power	1,219
GM	617	5.9	0.92	Power	781
CB	801	6.1	0.91	Power	996
Drainage system					
Churince	36,104	8.7			1,780
La Becerra	29,401	9.2			192
Pozas Rojas	4,217	7.4			1,538

cyanophage core proteins" were also found, but less abundantly (data not shown). Other abundant functional subsystems were (i) cofactors, vitamins, prosthetic groups, and pigments; (ii) DNA metabolism; (iii) nucleosides and nucleotides; (iv) regulation and cell signaling; and (v) RNA metabolism (Table 3). The functional categories were compared between aquatic subsystems with a PERMANOVA (a nonparametric multivariate permutation test); subsystems i, iv, and v were more significantly overrepresented in Churince (*P* values of 0.008, 0.000, and 0.000, respectively), and "Phages, prophages, transposable elements, plasmids" were overrepresented in Pozas Rojas (*P* value of 0.004). More reads were assigned for the InterPro2GO functional profile than for the SEED analysis (24.7% versus 71.2%) (Table 3). The GO systems more represented were "Transferase activity (0016740)," "Hydrolase activity (0016787)," and "Nucleotide binding (0000166)". Other less well represented categories were "DNA metabolic process (0006259)" and "Biosynthetic process (0009058)." The proteins with more hits in these GO systems were related to helicase, DNA polymerase, RNA polymerase, and terminase phage activities.

Geographic structure of CCB viruses. Viral community diversity estimations were performed using the Phage Communities from Contig Spectra (PHACCS) analysis tool and Circonspect to generate the contig spectrum of each sample (Table 4 and Materials and Methods). Overall, the Shannon diversity indexes (H′) showed that all metagenomes have high diversity, namely, between 7.6 and 9.1 for Churince, 6.1 and 6.9 for Pozas Rojas, 9.2 for La Becerra, and 5.9 and 6.1 for fish contents (Table 4). Churince and La Becerra are more diverse than Pozas Rojas, as measured by one-way analysis of variance (ANOVA; *F* value of 7.9 and *P* value of 0.02 for Churince and La Becerra and *F* value of 66.0 and *P* value of 0.001 for Pozas Rojas). This index takes into account the number of species and their equitability in a sample; the diversity index will increase by having a higher number of species or by having a greater evenness of these species. The high diversity of Churince and La Becerra is related to the number of species (genotypes) in each system, since they had significantly more genotypes (11,795 to 34,678) than Pozas Rojas (929 to 3,663) (Table 4). On the other hand, the water metagenomes showed high evenness values (0.77 to 0.89) (Table 4) without significant differences between water systems. The fish samples have lower Shannon values (5.9 to 6.1) and richness (617 to 801) than water samples, but their evenness values are close to one (0.91 to 0.93), meaning that all of their virus genotypes are close to being equally abundant. These observations, along with the rarefaction curves, which are all nearing

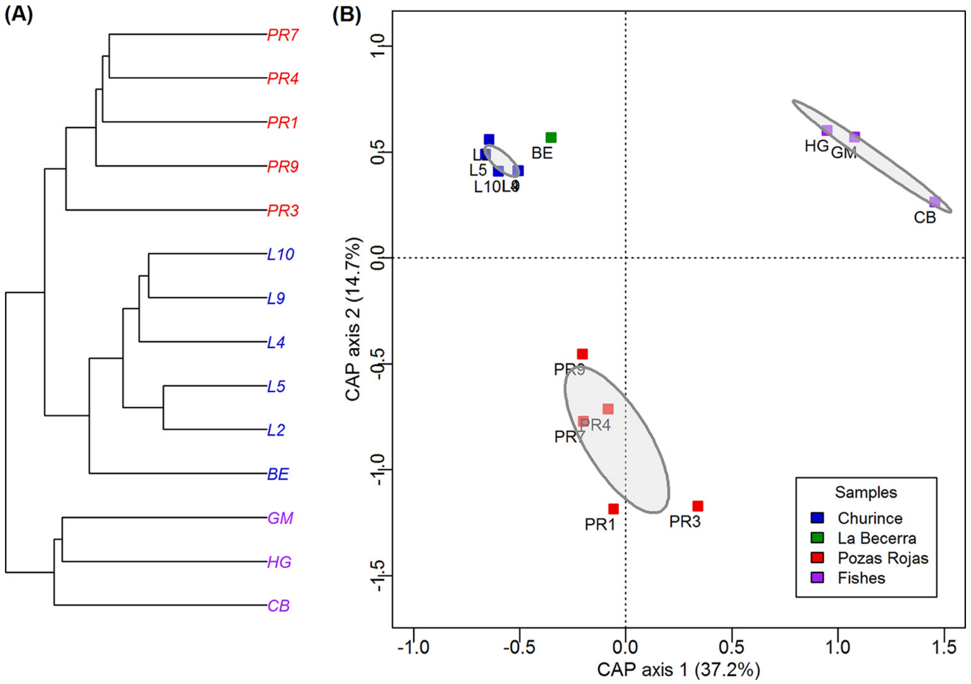


FIG 6 Canonical analysis of principal coordinates (PAC) and hierarchical clustering of viral communities at the species level. (A) Hierarchical clustering. (B) PAC analysis. Each point corresponds to a sample, ellipses represent the standard deviations of the weighted averages of the drainage systems. Ellipses were calculated using the Ordiellipse function of the R package vegan (59) at a 95% confidence level. HG, *Hemicromis gutatus*; GM, *Gambusia marshi*; CB, *Cyprinodon bifasciatus*.

an asymptote (Fig. S2), imply that our sampling effort within the area was adequate and that the estimates of diversity are close to the real value of viral diversity at the time of the survey.

With the canonical analysis of principal coordinates (PAC) based on Bray-Curtis dissimilarity distances from species abundance of each sample, normalized to smaller valid read samples, we examined the relations of the viral community between samples of CCB (Fig. 6B). Three significant clusters were observed. One cluster includes all of the samples from Churince and the sample from La Becerra, which are nearby systems (less than 5 km according to Google Earth) (Fig. 1). A second cluster includes the 5 samples from Pozas Rojas, on the other side of the valley (Fig. 1), where calcium carbonate dominates the chemistry of the water (34, 37). Finally, the third cluster includes the samples collected from the different species of fish within Churince, whose diet and physiology are enriched for particular viral communities (Fig. 6B). It is important to point out that PAC analysis also was done at the genus taxonomic and functional levels using SEED categories (level 2); both of them showed the same three geographical cluster patterns (data not shown).

To evaluate the robustness of this clustering pattern, we performed a hierarchical cluster analysis at the species level (Fig. 6A), confirming the confidence of the nodes and the similarities described in the previous representation. The significant difference between the three potential clusters was verified with PERMANOVA using Adonis (*F* value of 5.41 and *P* value of 0.006 for Churince, La Becerra, and Pozas Rojas; *F* value of 10.1 and *P* value of 0.022 for Churince, La Becerra, and FIC; and *F* value of 5.1 and *P* value of 0.018 for Pozas Rojas and FIC).

Phylogenetic analysis of the main viral families confirms endemism. The phylogenetic analyses confirmed the uniqueness of CCB; they revealed the biogeographic isolation signature in the 7 major viral families associated with different hosts that were used as examples. Different protein sequences, previously reported as markers for viral phylogeny analysis, were considered (1). The viral reads of these families were assem-

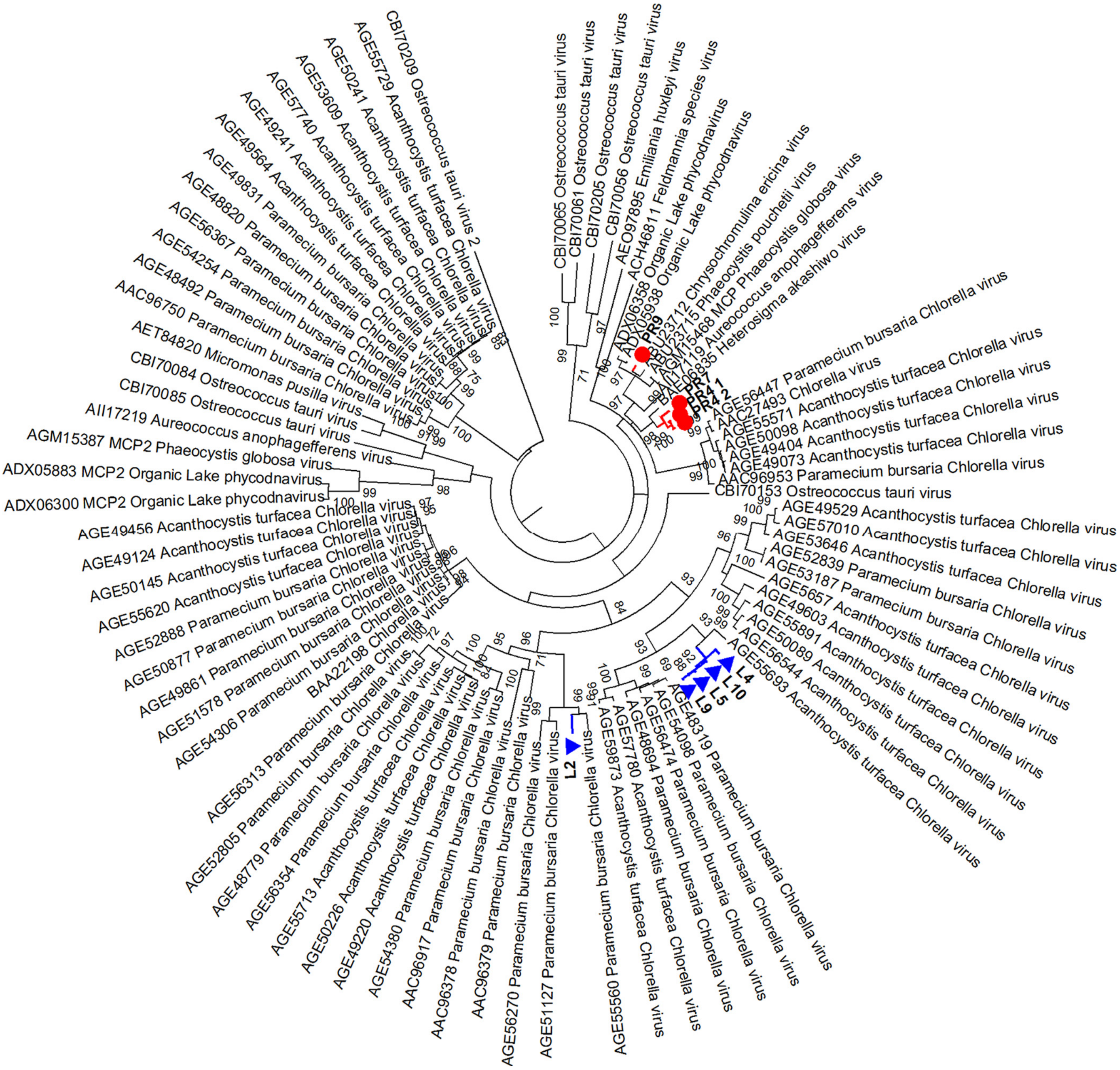

CBI70209 Ostreococcus tauri virus 2
AGE55729 Acanthocystis turfacea Chlorella virus
AGE50241 Acanthocystis turfacea Chlorella virus
AGE53609 Acanthocystis turfacea Chlorella virus
AGE57740 Acanthocystis turfacea Chlorella virus
AGE49241 Acanthocystis turfacea Chlorella virus
AGE49564 Acanthocystis turfacea Chlorella virus
AGE49831 Paramecium bursaria Chlorella virus
AGE48820 Paramecium bursaria Chlorella virus
AGE56367 Paramecium bursaria Chlorella virus
AGE54254 Paramecium bursaria Chlorella virus
AGE48492 Paramecium bursaria Chlorella virus
AAC96750 Paramecium bursaria Chlorella virus
AET84820 Micromonas pusilla virus
CBI70084 Ostreococcus tauri virus
CBI70085 Ostreococcus tauri virus
AII17219 Aureococcus anophagefferens virus
AGM15387 MCP2 Phaeocystis globosa virus
ADX05883 MCP2 Organic Lake phycodnavirus
ADX06300 MCP2 Organic Lake phycodnavirus
AGE49456 Acanthocystis turfacea Chlorella virus
AGE49124 Acanthocystis turfacea Chlorella virus
AGE50145 Acanthocystis turfacea Chlorella virus
AGE55620 Acanthocystis turfacea Chlorella virus
AGE52888 Paramecium bursaria Chlorella virus
AGE50877 Paramecium bursaria Chlorella virus
AGE49861 Paramecium bursaria Chlorella virus
AGE51578 Paramecium bursaria Chlorella virus
AGE54306 Paramecium bursaria Chlorella virus
BAA22198 Chlorella virus
AGE56313 Paramecium bursaria Chlorella virus
AGE52805 Paramecium bursaria Chlorella virus
AGE48779 Paramecium bursaria Chlorella virus
AGE56354 Paramecium bursaria Chlorella virus
AGE55713 Acanthocystis turfacea Chlorella virus
AGE50226 Acanthocystis turfacea Chlorella virus
AGE49220 Acanthocystis turfacea Chlorella virus
AGE54380 Paramecium bursaria Chlorella virus
AAC96917 Paramecium bursaria Chlorella virus
AAC96378 Paramecium bursaria Chlorella virus
AAC96379 Paramecium bursaria Chlorella virus
AGE56270 Paramecium bursaria Chlorella virus
AGE51127 Paramecium bursaria Chlorella virus
L2
AGE55560 Paramecium bursaria Chlorella virus
AGE59873 Acanthocystis turfacea Chlorella virus
AGE57780 Acanthocystis turfacea Chlorella virus
AGE48694 Paramecium bursaria Chlorella virus
AGE56474 Paramecium bursaria Chlorella virus
AGE54098 Paramecium bursaria Chlorella virus
AGE48319 Paramecium bursaria Chlorella virus
L9
L5
L10
L4
AGE55693 Acanthocystis turfacea Chlorella virus
AGE56544 Acanthocystis turfacea Chlorella virus
AGE50089 Acanthocystis turfacea Chlorella virus
AGE55891 Acanthocystis turfacea Chlorella virus
AGE49603 Acanthocystis turfacea Chlorella virus
AGE5657 Acanthocystis turfacea Chlorella virus
AGE53187 Paramecium bursaria Chlorella virus
AGE52839 Paramecium bursaria Chlorella virus
AGE53646 Acanthocystis turfacea Chlorella virus
AGE57010 Acanthocystis turfacea Chlorella virus
AGE49529 Acanthocystis turfacea Chlorella virus
CBI70153 Ostreococcus tauri virus
AAC96953 Paramecium bursaria Chlorella virus
AGE49073 Acanthocystis turfacea Chlorella virus
AGE49404 Acanthocystis turfacea Chlorella virus
AGE50098 Acanthocystis turfacea Chlorella virus
AGE55571 Acanthocystis turfacea Chlorella virus
AAC27493 Chlorella virus
AGE56447 Paramecium bursaria Chlorella virus
PR4 2
PR4 1
PR7
BAE06835 Heterosigma akashiwo virus
AII17119 Aureococcus anophagefferens virus
AGM15468 MCP Phaeocystis globosa virus
ABU23715 Phaeocystis pouchetii virus
ABU23712 Chrysochromulina ericina virus
PR9
ADX05938 Organic Lake phycodnavirus
ADX06358 Organic Lake phycodnavirus
ACH46811 Feldmannia species virus
AEO97895 Emiliania huxleyi virus
CBI70056 Ostreococcus tauri virus
CBI70205 Ostreococcus tauri virus
CBI70061 Ostreococcus tauri virus
CBI70065 Ostreococcus tauri virus

FIG 7 Maximum likelihood tree of the major capsid protein of the *Phycodnaviridae* family. In total, 254 sequences were downloaded from GenBank, and a subset of 82 was used for the analysis after eliminating sequences with 95% identity using CD-hit. Branches in the tree are colored according to the following system: blue for Churince contigs and red for Pozas Rojas. Bootstrap values are indicated for each node having at least 50% support.

bled into contigs long enough to build informative phylogenetic trees for the targeted viral groups (Table S3).

(i) For the dsDNA *Phycodnaviridae* family, which infects algae, the major capsid VP54, the ribonucleoside-triphosphate reductase, and the thymidylate synthase proteins were used for three different phylogenetic analyses. Based on the major capsid protein (Fig. 7), the contig sequences obtained from Churince mainly mapped to a new clade related to *Acanthocystis turfacea* chlorella virus. On the other hand, some contigs from Pozas Rojas exhibited more similarities to a different clade that includes the *Heterosigma akashiwo* virus. The analysis based on the ribonucleoside-triphosphate reductase protein showed that none of the contigs were affiliated with a known lineage in the tree (Fig. 8) and indicated the novelty of the CCB virome; indeed, all contigs were

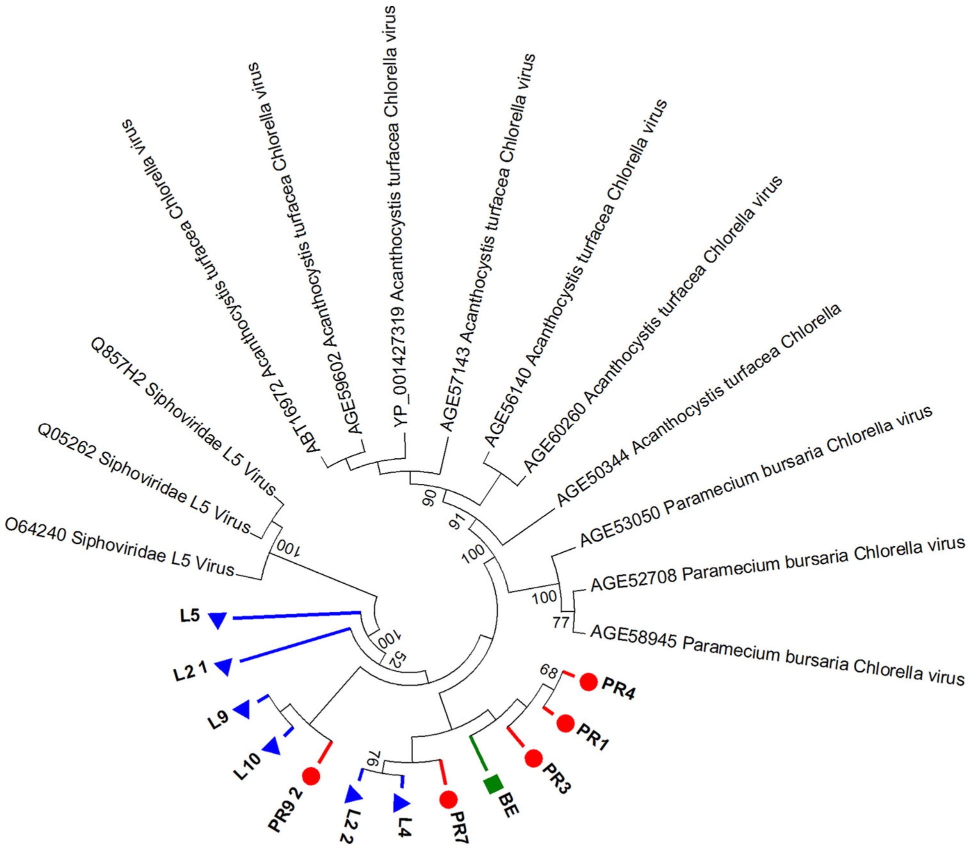

ABT16972 Acanthocystis turfacea Chlorella virus
AGE59602 Acanthocystis turfacea Chlorella virus
YP_001427319 Acanthocystis turfacea Chlorella virus
AGE57143 Acanthocystis turfacea Chlorella virus
AGE56140 Acanthocystis turfacea Chlorella virus
AGE60260 Acanthocystis turfacea Chlorella virus
AGE50344 Acanthocystis turfacea Chlorella
AGE53050 Paramecium bursaria Chlorella virus
AGE52708 Paramecium bursaria Chlorella virus
AGE58945 Paramecium bursaria Chlorella virus
Q857H2 Siphoviridae L5 Virus
Q05262 Siphoviridae L5 Virus
O64240 Siphoviridae L5 Virus
100
90
91
100
100
77
100
52
68
76
L5
L2 1
L9
L10
PR9 2
L2 2
L4
PR7
BE
PR3
PR1
PR4

FIG 8 Maximum likelihood tree of the ribonucleoside-triphosphate reductase protein of the *Phycodnaviridae* family. In total, 10 *Phycodnaviridae* sequences were used as the reference, and 3 *Shipoviridae* sequences were also included, since they were the next most homologous sequences in GenBank according to BLASTX. Branches in the tree are colored according to the following codes: blue for Churince contigs, red for Pozas Rojas, and green for La Becerra. Bootstrap values are indicated for each node having at least 50% support.

placed in a new clade close to *Phycodnaviridae* viruses. Finally, the thymidylate synthase protein analysis placed the contigs from Churince, La Becerra, and Pozas Rojas close to only one virus clade, the Organic Lake phycodnavirus 1 (data not shown).

(ii) The phylogenetic analysis of the plant-infecting *Nanoviridae* virus family (Fig. 9) placed some contigs from Churince and Pozas Rojas near the replication protein of *Picobiliphyte* sp. strain MS584-5 nanovirus lineage, but other contigs did not show any evident relationship to known viruses.

(iii) For the viruses in the *Iridoviridae* family, which infect mainly invertebrates but also frogs and fishes, the phylogeny was constructed using the ribonucleoside reductase alpha subunit protein. The analysis showed that in general, all contigs defined many novel lineages (Fig. 10). Moreover, PR3, PR4, and PR7 were placed in a new major clade. Of note, the Churince and Pozas Rojas contigs branched in different parts of the tree.

(iv) For viruses in the *Mimiviridae* family, which infects protozoa, the ATP-dependent RNA helicase and the *ankyrin* repeat proteins were used for phylogenetic analysis. Based on the RNA helicase proteins (Fig. 11), only one contig was affiliated with a known virus (*Acanthamoeba polyphaga* mimivirus), while the rest of them defined major novel branches without exhibiting any sequence relationship with the reference viruses. Finally, the *ankyrin* repeat protein phylogeny showed that contigs from all samples were placed with *Moumouvirus monve* and *Moumouvirus goulette* viruses (data no shown).

(v) Lastly, for the *Myoviridae*, *Siphoviridae*, and *Podoviridae* bacteriophage families, the phylogenetic analyses were done using the terminase large subunit (Fig. 12) and

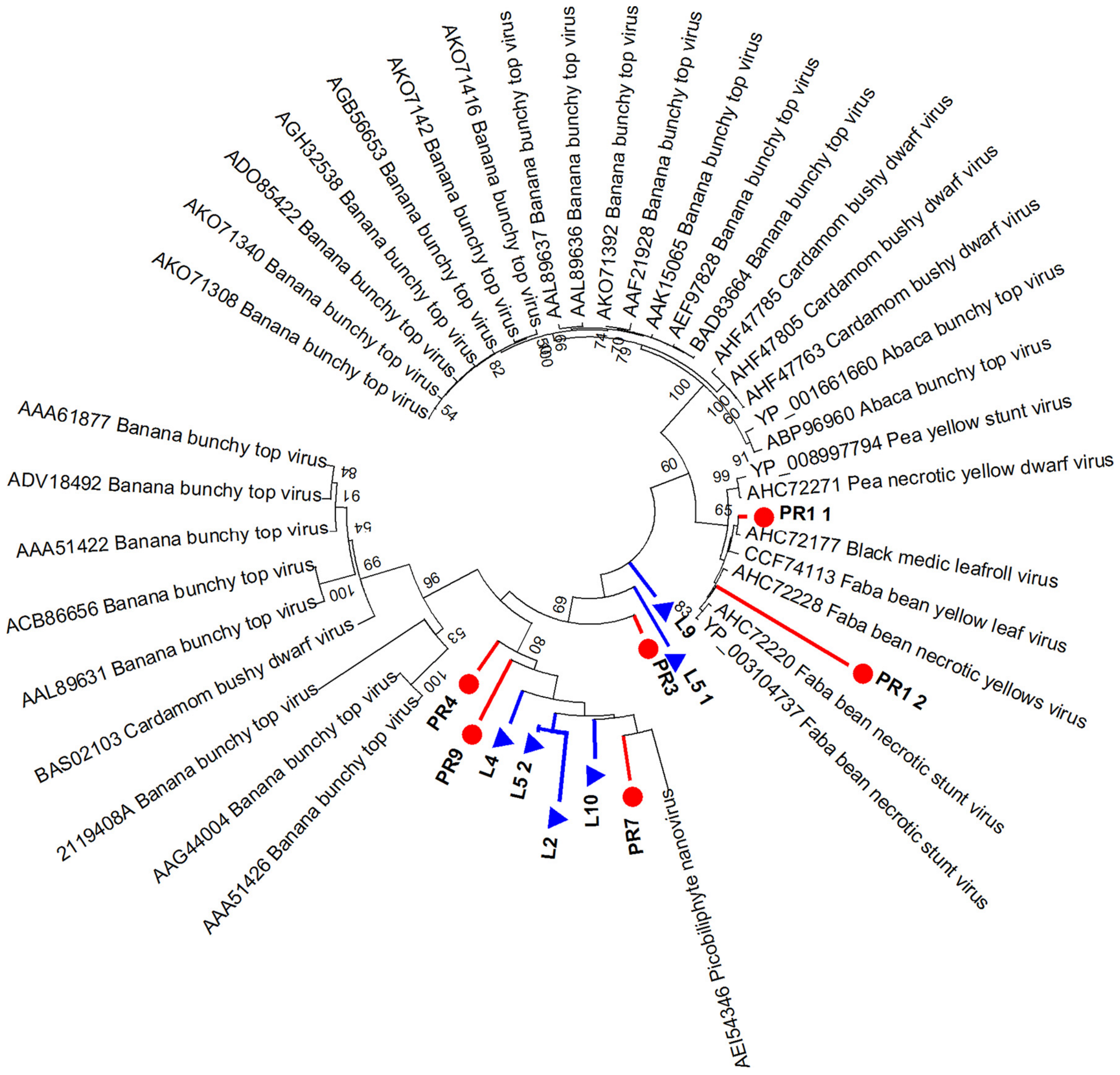

AKO71308 Banana bunchy top virus
AKO71340 Banana bunchy top virus
ADO85422 Banana bunchy top virus
AGH32538 Banana bunchy top virus
AGB56653 Banana bunchy top virus
AKO7142 Banana bunchy top virus
AKO71416 Banana bunchy top virus
AAL89637 Banana bunchy top virus
AAL89636 Banana bunchy top virus
AKO71392 Banana bunchy top virus
AAF21928 Banana bunchy top virus
AAK15065 Banana bunchy top virus
AEF97828 Banana bunchy top virus
BAD83664 Banana bunchy top virus
AHF47785 Cardamom bushy dwarf virus
AHF47805 Cardamom bushy dwarf virus
AHF47763 Cardamom bushy dwarf virus
YP_001661660 Abaca bunchy top virus
ABP96960 Abaca bunchy top virus
YP_008997794 Pea yellow stunt virus
AHC72271 Pea necrotic yellow dwarf virus
PR1 1
AHC72177 Black medic leafroll virus
CCF74113 Faba bean yellow leaf virus
AHC72228 Faba bean necrotic yellows virus
PR1 2
AHC72220 Faba bean necrotic stunt virus
YP_003104737 Faba bean necrotic stunt virus
AEI54346 Picobiliphyte nanovirus
PR7
L10
L2
L5 2
L4
PR9
PR4
PR3
L5 1
L9
AAA51426 Banana bunchy top virus
AAG44004 Banana bunchy top virus
2119408A Banana bunchy top virus
BAS02103 Cardamom bushy dwarf virus
AAL89631 Banana bunchy top virus
ACB86656 Banana bunchy top virus
AAA51422 Banana bunchy top virus
ADV18492 Banana bunchy top virus
AAA61877 Banana bunchy top virus
54
82
100
96
74
79
100
100
60
91
99
60
65
83
69
96
66
84
91
54
100
53
80
100

FIG 9 Maximum likelihood tree of the replication initiator protein of the *Nanoviridae* family. In total, 673 sequences were downloaded from GenBank, and a subset of 38 was used for the analysis after eliminating sequences with 97% identity using the CD-hit program. Branches in the tree are colored according to the following codes: blue for Churince contigs and red for Pozas Rojas. Bootstrap values are indicated for each node having at least 50% support.

the major capsid protein (Fig. 13). The phylogenies showed that contigs from all samples not only recaptured several known lineages of all tree viral families but also defined many novel major branches, depending on the reference protein used.

DISCUSSION

Viral diversity is related to ecosystem diversity and structure. In this study, we described the general virus *inventarium* of Churince at CCB, comparing the viral diversity of this very endangered site with those of La Becerra and Pozas Rojas, two sites under recovery after different types of perturbations. We also determined the virus composition filtered by the fish diet in the guts of fishes from Churince. Our analysis included RNA and DNA viruses and covered different collection sites, with much larger sequence coverage than a previous study reporting viruses associated with the microbialites in CCB (35).

Among the viral sequences detected, there was a large inequality between DNA (75.4%) and RNA (0.5%) viruses (Fig. 2); this may be related to a true lower abundance of RNA viruses, smaller genomes sizes, or the fact that RNA is more prone to degradation than DNA.

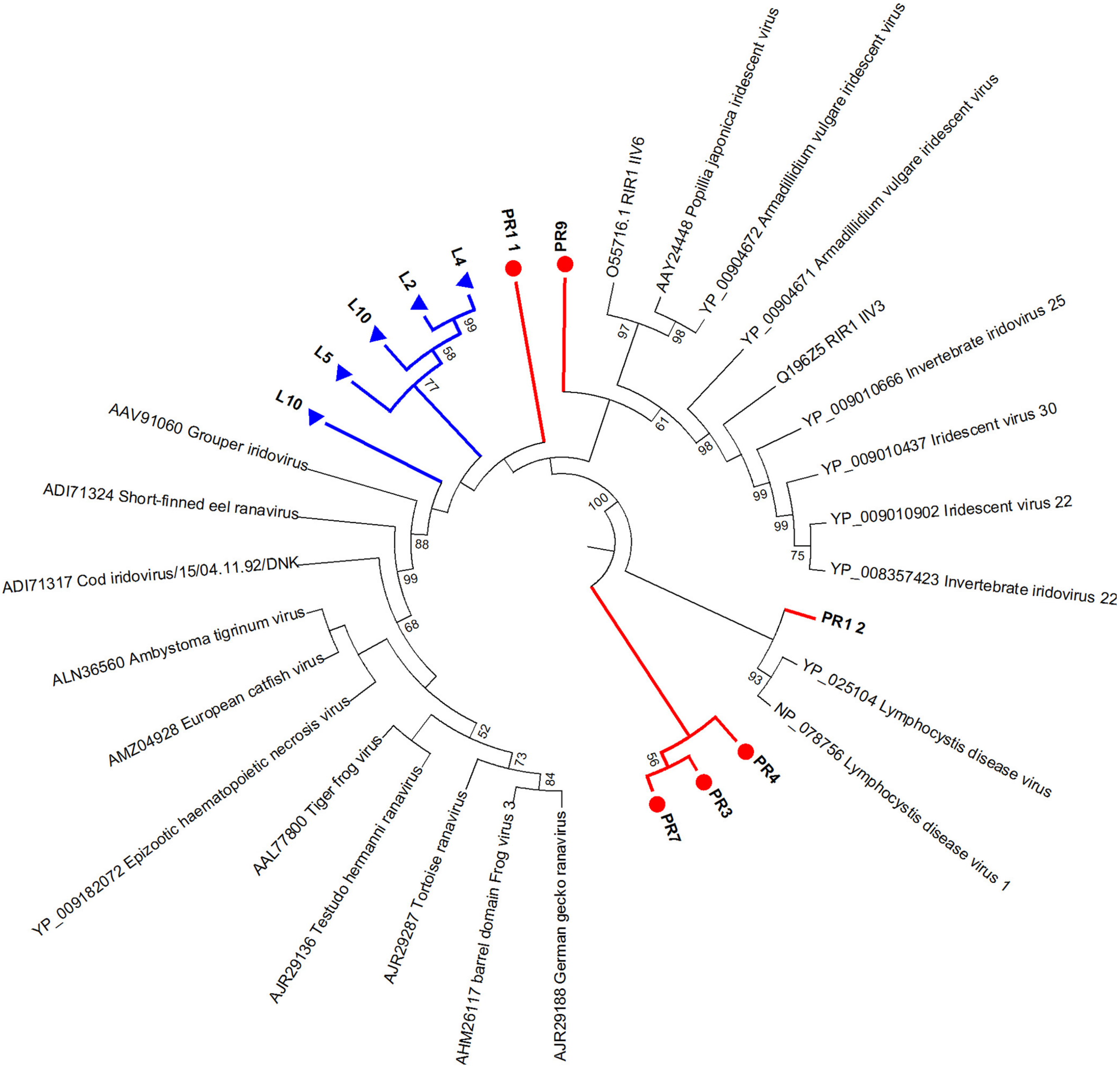

O55716.1 RIR1 IIV6
AAY24448 Popillia japonica iridescent virus
YP_00904672 Armadillidium vulgare iridescent virus
YP_00904671 Armadillidium vulgare iridescent virus
Q196Z5 RIR1 IIV3
YP_009010666 Invertebrate iridovirus 25
YP_009010437 Iridescent virus 30
YP_009010902 Iridescent virus 22
YP_008357423 Invertebrate iridovirus 22
PR1 2
YP_025104 Lymphocystis disease virus
NP_078756 Lymphocystis disease virus 1
PR4
PR3
PR7
AJR29188 German gecko ranavirus
AHM26117 barrel domain Frog virus 3
AJR29287 Tortoise ranavirus
AJR29136 Testudo hermanni ranavirus
AAL77800 Tiger frog virus
YP_009182072 Epizootic haematopoietic necrosis virus
AMZ04928 European catfish virus
ALN36560 Ambystoma tigrinum virus
ADI71317 Cod iridovirus/15/04.11.92/DNK
ADI71324 Short-finned eel ranavirus
AAV91060 Grouper iridovirus
L10
L5
L10
L2
L4
PR1 1
PR9
97
98
61
98
99
99
75
100
93
56
99
58
77
88
99
68
52
73
84

FIG 10 Maximum likelihood tree of the ribonucleoside reductase alpha subunit protein of the *Iridoviridae* family. In total, 79 sequences from this protein were downloaded from GenBank and 18 were used in the analysis after a CD-hit at 99% similarity. Branches in the tree are colored according to the following codes: blue for Churince contigs and red for Pozas Rojas. Bootstrap values are indicated for each node having at least 50% support.

In this regard, the few studies that have addressed the presence of RNA viruses in the ocean also found lower abundances of RNA viruses than DNA viruses (2).

Between 63.1% and 82.2% of the sequence reads showed no identity to any sequences in the searched databases (see Fig. S1 in the supplemental material). A similar percentage has been reported in ocean samples (5–14). Even in the latest ocean survey (TARA), the amount of what was considered viral dark matter ranged from 63 to 93% (2). It is possible that some of our unknown sequences correspond to uncharacterized viruses that prey on the diverse and divergent bacteria found in CCB, since bacteria are the most diverse part of the aquatic communities in CCB (38).

Of interest, after bacteriophages and invertebrate- and plant-related viruses, the next most abundant sequence reads, which were present in all samples, shared identity with

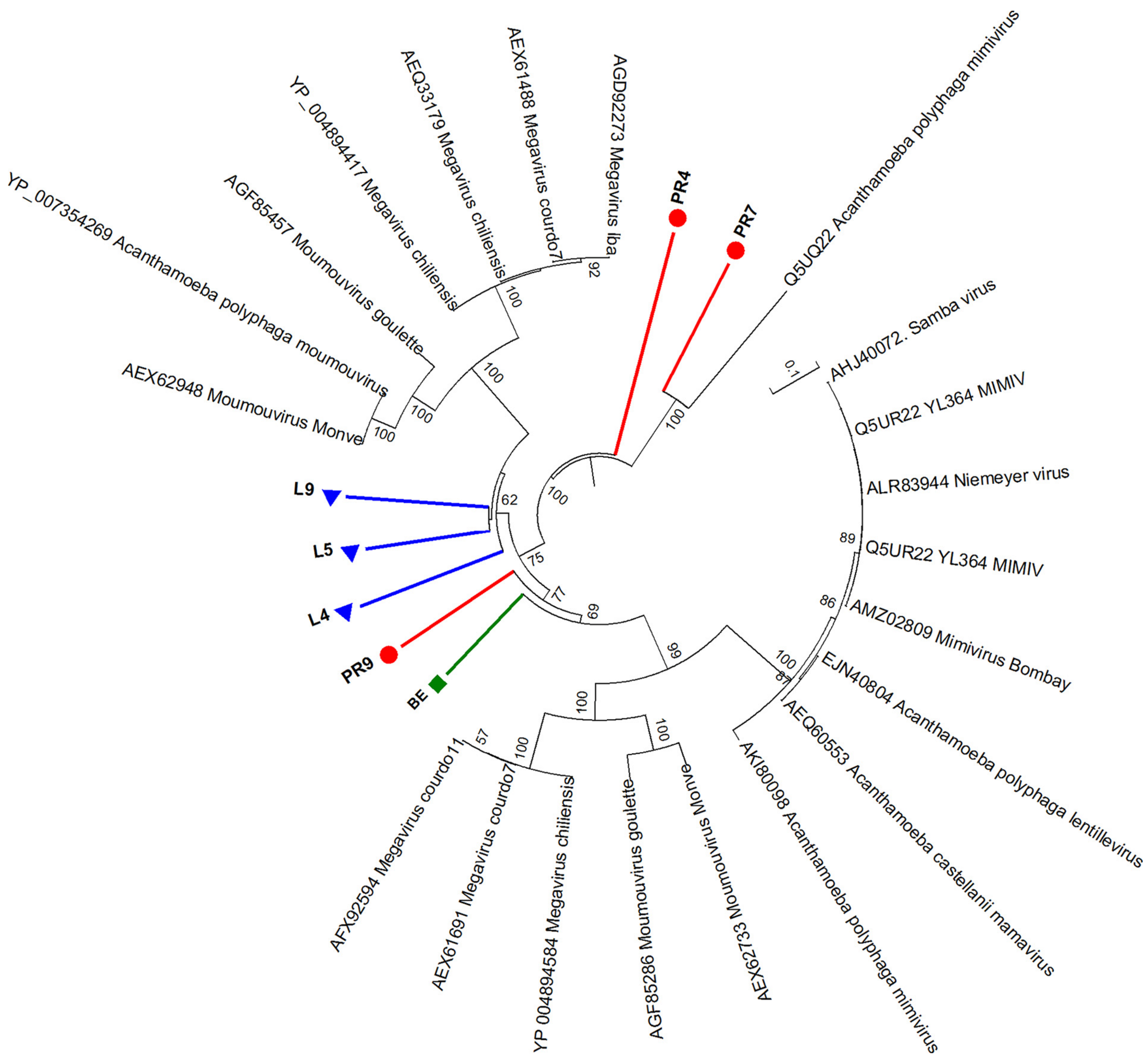

AGD92273 Megavirus iba
AEX61488 Megavirus courdo7
AEQ33179 Megavirus chiliensis
YP_004894417 Megavirus chiliensis
AGF85457 Moumouvirus goulette
YP_007354269 Acanthamoeba polyphaga moumouvirus
AEX62948 Moumouvirus Monve
PR4
PR7
Q5UQ22 Acanthamoeba polyphaga mimivirus
0.1
AHJ40072. Samba virus
Q5UR22 YL364 MIMIV
ALR83944 Niemeyer virus
Q5UR22 YL364 MIMIV
AMZ02809 Mimivirus Bombay
EJN40804 Acanthamoeba polyphaga lentillevirus
AEQ60553 Acanthamoeba castellanii mamavirus
AKI80098 Acanthamoeba polyphaga mimivirus
AEX62733 Moumouvirus Monve
AGF85286 Moumouvirus goulette
YP 004894584 Megavirus chiliensis
AEX61691 Megavirus courdo7
AFX92594 Megavirus courdo11
L9
L5
L4
PR9
BE
92
100
100
100
100
100
100
62
75
77
69
99
100
100
57
100
89
86
100
87

FIG 11 Maximum likelihood tree of the ATP-dependent RNA helicase protein of *Mimiviridae* family. In total, 21 sequences were used as a reference. Branches in the tree are colored according to the following codes: blue for Churince contigs, red for Pozas Rojas, and green from Becerra. Bootstrap values are indicated for each node having at least 50% support.

eukaryotic dsDNA virus families, including the *Phycodnaviridae* (0.9% to 5.7%) (Fig. 3), whose members are known to infect algae. This was surprising given the ultraoligotrophic nature of the site, where no visible algae are observed, although microalgae are presumably present. On the other hand, recent evidence has shown that phycodnaviruses can successfully infect nonalgae hosts, including humans and mice (39, 40). These observations reinforce the hypothesis that these viruses, as well as other giant viruses infecting amoebas or other unicellular protists, cause opportunistic infections, as previously reported (41, 42). Viruses that infect free-living amoebae, from the family *Mimiviridae*, were also found in all samples, suggesting that the protozoan biodiversity is also large, an observation confirmed by 18S rRNA gene libraries (38). The amino acid identities to known viral proteins were less than 60% in most cases, suggesting that most of these viruses are novel members of identified virus families. It is worth mentioning that extraction controls were not included, so some of the detected viruses could be contaminants from the extraction process.

The unique and extensive virome diversity of CCB suggested by the sequence identity data was confirmed by the phylogenetic analyses. The phylogeny showed that contigs from all samples recaptured several known lineages but also defined many

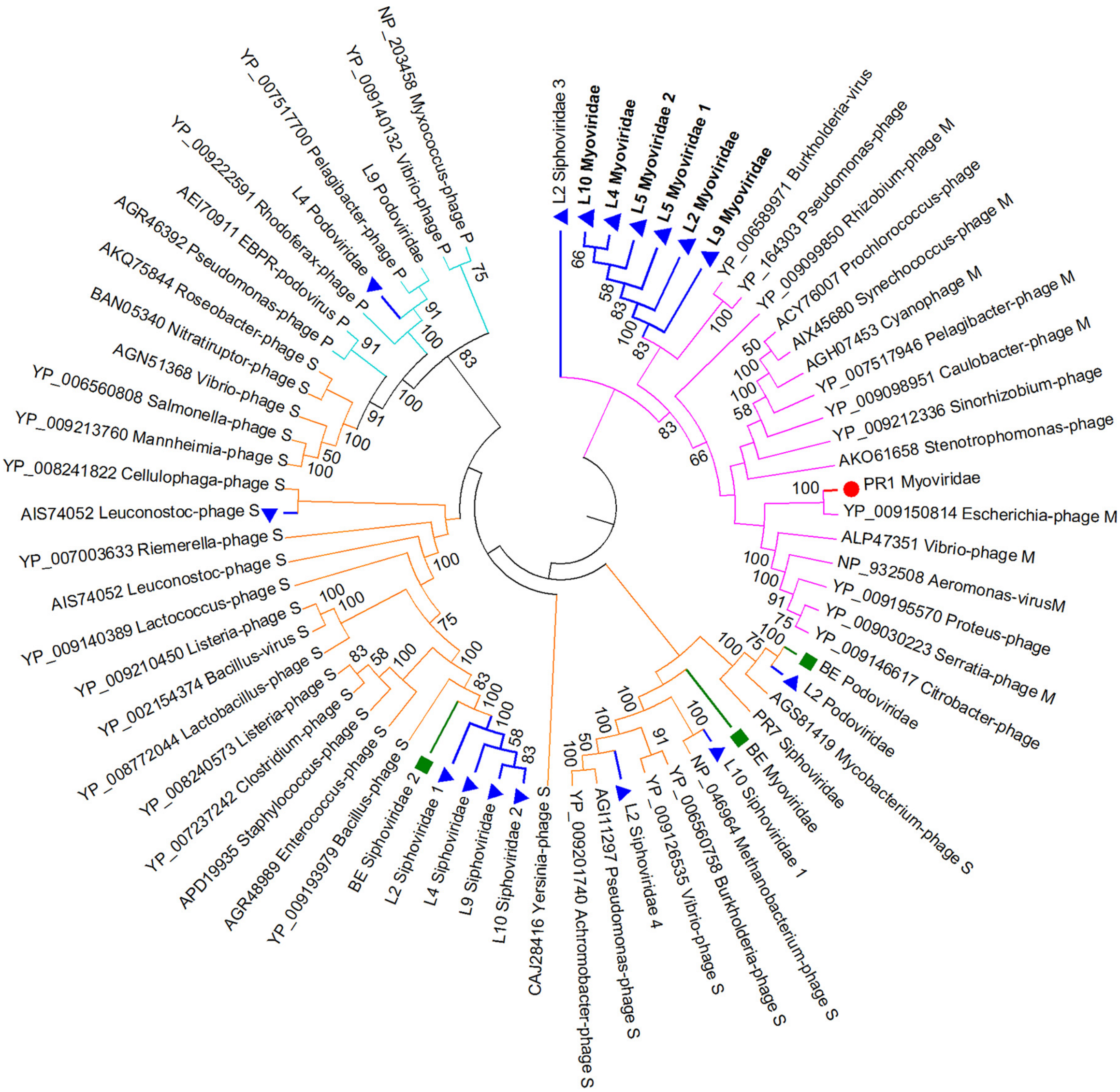

NP_203458 Myxococcus-phage P
YP_009140132 Vibrio-phage P
L9 Podoviridae
YP_007517700 Pelagibacter-phage P
L4 Podoviridae
YP_009222591 Rhodoferax-phage P
AEI70911 EBPR-podovirus P
AGR46392 Pseudomonas-phage S
AKQ75844 Roseobacter-phage S
BAN05340 Nitratiruptor-phage S
AGN51368 Vibrio-phage S
YP_006560808 Salmonella-phage S
YP_009213760 Mannheimia-phage S
YP_008241822 Cellulophaga-phage S
AIS74052 Leuconostoc-phage S
YP_007003633 Riemerella-phage S
AIS74052 Leuconostoc-phage S
YP_009140389 Lactococcus-phage S
YP_009210450 Listeria-phage S
YP_002154374 Bacillus-virus S
YP_008772044 Lactobacillus-phage S
YP_008240573 Listeria-phage S
YP_007237242 Clostridium-phage S
APD19935 Staphylococcus-phage S
AGR48989 Enterococcus-phage S
YP_009193979 Bacillus-phage S
BE Siphoviridae 2
L2 Siphoviridae 1
L4 Siphoviridae
L9 Siphoviridae
L10 Siphoviridae 2
CAJ28416 Yersinia-phage S
YP_009201740 Achromobacter-phage S
AGI11297 Pseudomonas-phage S
L2 Siphoviridae 4
YP_009126535 Vibrio-phage S
YP_006560758 Burkholderia-phage S
NP_046964 Methanobacterium-phage S
L10 Siphoviridae 1
BE Myoviridae
PR7 Siphoviridae
AGS81419 Mycobacterium-phage S
L2 Podoviridae
BE Podoviridae
YP_009146617 Citrobacter-phage
YP_009030223 Serratia-phage M
YP_009195570 Proteus-phage
NP_932508 Aeromonas-virusM
ALP47351 Vibrio-phage M
YP_009150814 Escherichia-phage M
PR1 Myoviridae
AKO61658 Stenotrophomonas-phage
YP_009212336 Sinorhizobium-phage
YP_009098951 Caulobacter-phage M
YP_007517946 Pelagibacter-phage M
AGH07453 Cyanophage M
AIX45680 Synechococcus-phage M
ACY76007 Prochlorococcus-phage
YP_009099850 Rhizobium-phage M
YP_164303 Pseudomonas-phage
YP_006589971 Burkholderia-virus
L9 Myoviridae
L2 Myoviridae
L5 Myoviridae 1
L5 Myoviridae 2
L4 Myoviridae
L10 Myoviridae
L2 Siphoviridae 3
75
91
100
83
91
100
91
100
50
100
100
100
75
100
100
100
83
58
100
100
83
100
100
58
83
100
50
100
91
100
100
75
100
100
91
100
100
100
66
83
58
83
100
83
66
100
50
100
100
58
100

FIG 12 Maximum likelihood tree of the terminase large subunit protein of the *Siphoviridae*, *Podoviridae*, and *Myoviridae* families that infect bacteria. Branches in the tree are colored according to the following codes: blue for Churince contigs, red for Pozas Rojas, and green from La Becerra. Internal branches are colored to identify families using the following codes: orange for *Siphoviridae*, pink for *Myoviridae*, and cyan for *Podoviridae*. Bootstrap values are indicated for each node having at least 50% support.

novel major branches, depending on the reference protein used. For example, the phylogeny analysis of the ribonucleoside-triphosphate reductase sequence of dsDNA viruses of the *Phycodnaviridae* family showed that none of the contigs were related to any previously known clade in the tree (Fig. 8). A similar phylogenetic pattern was found for the viruses in the plant-infecting family *Nanoviridae* (Fig. 9). The phylogeny showed that some virus proteins from Churince and Pozas Rojas were placed in the same clade of *Picobiliphyte* sp. strain MS584-5 nanovirus, while other contigs did not show an evident relation to any known virus.

Regarding the viral diversity found in the FIC, the insect-related dicistroviruses were dominant (91%) in the endangered microendemic pupfish (*C. bifasciatus*); these viruses were also abundant in other fishes but in a lesser proportion. The most abundant virus families in the invasive fish are related to bacteria. At variance with viruses in the water samples, the most abundant phages species found in the FIC did not show similarity to viruses isolated from a marine environment (Table S2).

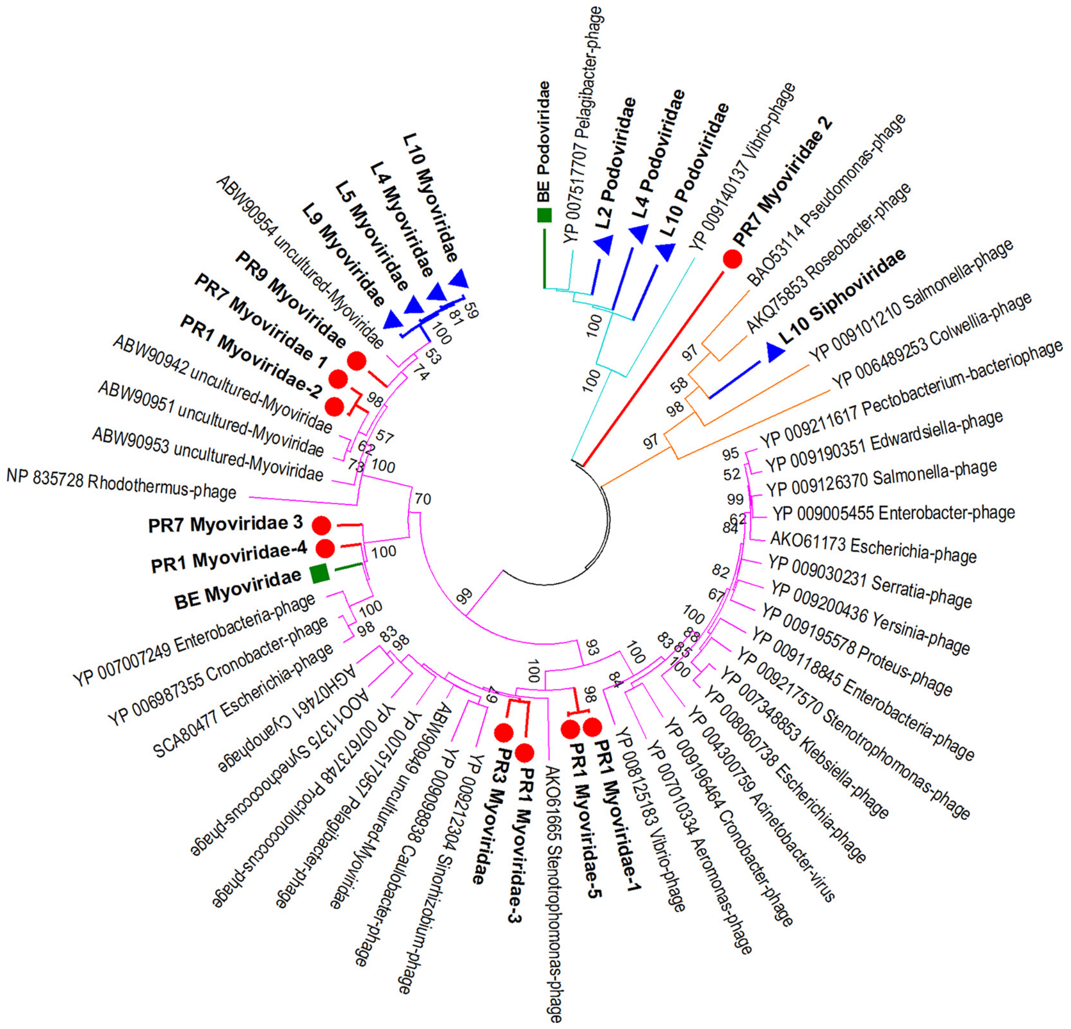

BE Podoviridae
YP 007517707 Pelagibacter-phage
L2 Podoviridae
L4 Podoviridae
L10 Podoviridae
YP 009140137 Vibrio-phage
PR7 Myoviridae 2
BAO53114 Pseudomonas-phage
AKQ75853 Roseobacter-phage
L10 Siphoviridae
YP 009101210 Salmonella-phage
YP 006489253 Colwellia-phage
YP 009211617 Pectobacterium-bacteriophage
YP 009190351 Edwardsiella-phage
YP 009126370 Salmonella-phage
YP 009005455 Enterobacter-phage
AKO61173 Escherichia-phage
YP 009030231 Serratia-phage
YP 009200436 Yersinia-phage
YP 009195578 Proteus-phage
YP 009118845 Enterobacteria-phage
YP 009217570 Stenotrophomonas-phage
YP 007348853 Klebsiella-phage
YP 008060738 Escherichia-phage
YP 004300759 Acinetobacter-virus
YP 009196464 Cronobacter-phage
YP 007010334 Aeromonas-phage
YP 008125183 Vibrio-phage
PR1 Myoviridae-1
PR1 Myoviridae-5
AKO61665 Stenotrophomonas-phage
PR1 Myoviridae-3
PR3 Myoviridae
YP 009212304 Sinorhizobium-phage
YP 009098938 Caulobacter-phage
ABW90949 uncultured-Myoviridae
YP 007517957 Pelagibacter-phage
YP 007673748 Prochlorococcus-phage
AOO11375 Synechococcus-phage
AGH07461 Cyanophage
SCA80477 Escherichia-phage
YP 006987355 Cronobacter-phage
YP 007007249 Enterobacteria-phage
BE Myoviridae
PR1 Myoviridae-4
PR7 Myoviridae 3
NP 835728 Rhodothermus-phage
ABW90953 uncultured-Myoviridae
ABW90951 uncultured-Myoviridae
ABW90942 uncultured-Myoviridae
PR1 Myoviridae-2
PR7 Myoviridae 1
PR9 Myoviridae
ABW90954 uncultured-Myoviridae
L9 Myoviridae
L5 Myoviridae
L4 Myoviridae
L10 Myoviridae
59
81
100
53
74
98
57
62
73
100
70
100
99
100
98
83
98
49
100
93
98
84
100
83
85
88
100
100
67
82
84
62
99
52
95
97
98
58
97
100
100

FIG 13 Maximum likelihood tree of major capsid proteins of *Siphoviridae*, *Podoviridae*, and *Myoviridae* families. Branches in the tree are colored according to the following codes: blue for Churince contigs, red for Pozas Rojas, and green from La Becerra. Internal branches are colored to identify families using the following codes: orange for *Siphoviridae*, pink for *Myoviridae*, and cyan for *Podoviridae*. Bootstrap is indicated for each node having at least 50% support.

Viral geographic structure, diversity, and relative abundances. In total, we detected 1,691 different viral species from 170 genera and 40 different viral families in all water and fish samples. The current viral taxonomic classification (43) comprises 104 different viral families, 505 viral genera, and 3,186 viral species; thus, our results correspond to more than half (53.8%) of the viral species and to one-third (38.5%) of the virus families known so far.

The viruses in the sampled sites were found to be varied as well as biodiverse (Table 4), confirming the initial observations in microbialites that CCB has divergent and unique viruses (35). This is probably due to the endemicity and uniqueness of the hosts. Regarding the most abundant virus species (Table S2), it was interesting to find that the spring in La Becerra had no species overrepresented, while both of the fluctuating systems (Churince and Pozas Rojas) had dominant but different virus species, isolated from seawater, as indicated by their distinct "clouds" in the canonical analysis of principal coordinates (Fig. 6). This was also confirmed at the genus taxonomic level and functional level using SEED categories (level 2), showing they have the same geographical distinct patterns (data not shown).

The study sites and the water samples collected in this study were different from the samples analyzed in a previous study in CCB; in the work by Desnues et al. (35) (Table 5),

TABLE 5 Comparison of the data reported in this publication with published viromes from other freshwater and marine water environments

Reference	Sample	Yr	Country/			Habitat	Depth	Sequence	No. of:		
or sample	designation	collected	region	Geographic zone	Type	type[a]	(m)	length	Sequences	Genotypes	H′
6	SAR	2005	Bahamas	Sargasso Sea	454	1	80	100	397,939	5,140	7.7
6	GOM	1996	Mexico	Gulf of Mexico	454	1	164	100	262,501	15,400	8.2
6	BBC	1996	Canada	British Columbia	454	1	245	100	414,964	129,000	10.8
6	Artic	2002	Canada	Arctic Sea	454	1	3246	100	686,209	532	6.05
10	GS112	2005	Seychelles	Indian Ocean	454	1	8	450	492,396	10,592	7.8
10	GS108	2005	Australia	Coccos Keeling	454	1	1.8	450	318,094	3,327	6.9
10	GS117	2005	Seychelles	St. Ann Island	454	1	1.8	450	710,101	22,040	8.8
10	GS122	2005	Australia	Between Madagascar/ South Africa	454	1	1.9	450	337,266	554	5.6
17	El Berbera	2012	Sahara	Mauritania	454	3	0	250	75,921	977	4.2
17	Hamdoun	2012	Sahara	Mauritania	454	3	0	220	39,404	91	2.2
17	Ilij	2012	Sahara	Mauritania	454	3	0	250	62,059	313	4.8
17	Molomhar	2012	Sahara	Mauritania	454	3	0	240	75,709	127	4.3
35	Pozas Azules	2005	Mexico	Pozas Azules II	454	2	0	100	301,264	19,520	8.9
35	Rio Mesquites	2005	Mexico	Rio Mesquites	454	2	0	100	324,500	33	3.0
35	Highborne Cay	2005	Bahamas	Highborne Cay	454	2	0	100	148,334	161	4.1
45	A	2011	USA	Lake Michigan	Illumina	3	10	250	4,404,942	1,609	6.9
45	F	2011	USA	Lake Michigan	Illumina	3	1	250	3,709,054	7,138	7.7
46	L.Spr.I.10 m	2009	Canada	Pacific Ocean	454	1	10	280	92,415	1,738	6.8
46	L.Sum.O.10 m	2009	Canada	Pacific Ocean	454	1	10	320	165,256	3,011	7.1
46	L.Win.O.10 m	2009	Canada	Pacific Ocean	454	1	10	290	192,685	4,961	7.8
46	L.Spr.O.10 m	2009	Canada	Pacific Ocean	454	1	10	280	75,036	1,506	6.9
46	L.Spr.C.10 m	2009	Canada	Pacific Ocean	454	1	10	270	107,244	10,355	8.1
46	M.Fall.O.10 m	2009	USA	Pacific Ocean	454	1	10	275	203,238	2,817	7.5
46	M.Fall.I.10 m	2009	USA	Pacific Ocean	454	1	10	300	321,754	30,470	9.0
46	M.Fall.C.10 m	2009	USA	Pacific Ocean	454	1	10	440	303,519	51,968	8.8
63	Spring	2011	Livingston Island	Antarctic Limnopolar Lake	454	3	10	220	41,322	5,111	6.5
63	Summer	2011	Livingston Island	Antarctic Limnopolar Lake	454	3	1	220	38,475	1,296	6.4
47	Lake Pavin	2007	France	Lake Pavin	454	3	20	400	649,290	587	4.74
47	Lake Bourget	2007	France	Lake Bourget	454	3	20	400	593,084	1,296	6.4
L10	L10	2014	Mexico	Churince	Illumina	3	0	130	6,137,055	15,302	7.4
L2	L2	2014	Mexico	Churince	Illumina	3	0	130	7,505,587	34,678	9.1
L4	L4	2014	Mexico	Churince	Illumina	3	0	130	8,251,324	11,795	7.6
L5	L5	2014	Mexico	Churince	Illumina	3	0	130	7,456,526	2,09044	8.6
L9	L9	2014	Mexico	Churince	Illumina	3	0	130	3,853,484	20,000	8.2
BE	BE	2014	Mexico	La Becerra	Illumina	3	0	130	6,973,697	24,901	9.2
PR1	PR1	2014	Mexico	Pozas Rojas	Illumina	3	0	130	4,353,067	2,765	6.9
PR3	PR3	2014	Mexico	Pozas Rojas	Illumina	3	0	130	4,247,703	929	6.1
PR4	PR4	2014	Mexico	Pozas Rojas	Illumina	3	0	130	10,359,155	2,235	6.7
PR7	PR7	2014	Mexico	Pozas Rojas	Illumina	3	0	130	6,133,145	3,663	6.6
PR9	PR9	2014	Mexico	Pozas Rojas	Illumina	3	0	130	5,461,242	3,314	6.5

[a] Habitat types were the following: 1, seawater; 2, microbialites; 3, freshwater.

two different sites were explored, a stromatolite of Pozas Azules and a trombolite in Rio Mezquites, which represent two of the most stable systems within the valley. In our study, the highest diversity corresponded to La Becerra water (H′, 9.2; evenness, 0.89), showing recovery of the ecosystem after years from anthropogenic disturbance due to tourism. Surprisingly, in Churince a high diversity (H′, 7.4 to 9.1), superior to that observed in the much less perturbed and seasonally fluctuating site of Pozas Rojas (H′, 6.1 to 6.9), was observed.

In relation to the FIC samples, the PAC and hierarchical clustering analysis showed that they were allocated in a group completely different from that of water samples, since they include a very particular distribution of the viral biodiversity. However, the virus communities in the jewelfish (*H. guttatus*) and in *Gambusia marshi* were more similar than the one in pupfish (*C. bifasciatus*), an observation that is also supported by their higher virus diversity compared to that of pupfish, possibly because either their feeding sources are broader or they inhabit a larger number of microniches (44).

Comparison to other freshwater and marine viral metagenomics data sets. We further evaluated CCB water viromes by comparing them to available data sets of other reported viral metagenomic studies. These studies include freshwater (15, 17, 45), marine water (6, 10, 46), and microbialites (35) and are summarized in Table 5. We compared our results by using the number of different genotypes and the Shannon diversity index from the PHACCS results reported in those works or by estimating them if they were not available, employing the same parameters as those used for our samples.

Second, Bray-Curtis distances of taxonomic species profiles were calculated, using the BLASTx classification results obtained from Metavir, and were used to represent the relative dissimilarity between samples.

The virome studies that analyzed ocean water in single samples have reported diversity Shannon indexes that ranged from 5.6 in the water between Madagascar and the Indian Ocean (10) to 9.0 in the North Pacific Ocean (United States) (46). In freshwater studies, Fancello et al. (17) characterized several ponds in a Mauritanian oasis in the Sahara desert and found a lower diversity that ranged from 2.2 to 4.8 (Table 5). This is in accordance with the diversity found in an isolated site in the French Lake Pavin, where the diversity was 4.7 in summer, compared with the extreme environmental conditions of an Antarctic lake, which showed a higher richness (6.4), or the human-impacted Lake Bourget and Michigan shores (45) that had higher diversities (6.4 to 7.7). Desnues et al. (35) showed that Pozas Azules stromatolite in CCB had a viral diversity of 8.9, while the stromatolite in the Bahamas had a lower diversity of 3.4. As mentioned above, in our study the Shannon index ranged from 6.1 to 6.9 in Pozas Rojas to 7.4 to 9.1 at Churince, showing a high diversity compared with other freshwater sites and even higher than or equal to that of some ocean samples (Table 5).

On the other hand, pairwise taxonomic comparison with these water samples, except for two old samples (6), were performed using Metavir results at the species level and Bray-Curtis dissimilarity distances (see Materials and Methods) that go from 0 (equal) to 1 (completely distinct). CCB water samples presented high Shannon diversity indexes, similar to those from seawater samples, and most of their abundant phages showed similarity to phages isolated from these marine environments; thus, the CCB samples were more similar to the Pacific Ocean (46) and Indian Ocean (10) samples than to freshwater samples (Table S4). The only exception was for the samples from Lake Pavin and Lake Burget (47), which were also similar to CCB and to the ocean samples mentioned above, even though Roux et al. (47) reported that viral communities of Lake Pavin and Lake Burget were more similar to the freshwater samples than to seawater. This discrepancy could be due to the fact that Roux et al. did not compare samples to a reference database but only to each other and used only a subsample of 50,000 reads. This dissimilarity was also presented when the PHACCS tool was used with 50,000 reads without resampling, which generated a large error in their abundance model.

Ocean viruses have been sampled carefully, and the observation that temperature as well as host abundance seemed to be more important than the geographic location for both viral abundance and diversity suggests that the geographical structure is not important (2, 48). The same observation was made in the Sahara study (17) and for the U.S. lake viromes (15, 45), where the season was more important than geography. On the other hand, a study of the viruses within *Sulfolobus islandicus*, a thermo acidophilic archaeon associated with thermal lakes in Russia, found clear signatures that viral biogeography follows the host biogeographic patterns (64). A similar viral biogeography was observed in microbialites in CCB, finding marine affinities and high divergence among each of the analyzed microbialite viral communities, as well as with the rest of the world, with the very important difference that in CCB, the virus geographical structure correlates with the high host beta diversity found mostly in bacteria (26).

Models explaining the diversity. Three main theories have dominated the literature regarding bacteriophage life strategies. The "kill the winner" model predicts that phages enter the lytic phase by density-dependent signals triggered when a host is very abundant, allowing other less abundant hosts to thrive for a while, thus increasing virus density and host biodiversity (48). The result of this pattern is an equitable distribution of species in a community over dominance (the dominant species gets killed). On the other hand, the "king of the hill" model suggests that the density of hosts is not related to the life cycle of the bacteriophages but is intrinsic to the virus lineage, and there is no clear relationship between virus abundance and host cell abundance, hence the dominant lineage can stay dominant. Finally, a third theory, called "piggyback on the winner," suggests that the most abundant host is invaded by lysogenic phages that ride into this successful, high-density host, hiding their genomes until the time is ripe; hence, the more cells, the less free virus is found (49). In this scenario, high dominance and lower diversity are expected.

The high virus biodiversity as well as equitable or close to equitable distributions found in our study suggest that the CCB systems follow a kill-the-winner model, with a high geographic differentiation (26) and a high microbial diversity (25–27, 34). That this is the case is suggested by the very high diversity, with low dominance and uniqueness found in each sampled CCB site, which represents a proportion of the hosts in the ecosystem and not just a few dominant well-adapted taxa.

We propose that in CCB, viruses are closely following their prey and evolving together, thereby explaining why both the hosts and the viruses have kept their ancient marine affiliations (27, 29, 32). This is supported by the high diversity found in CCB samples, similar to those of oceans, the more taxonomical similarity to viruses found in samples of marine origins than those found in freshwater, and finally to the finding that all of the more abundant phages found in CCB have been isolated from seawater. However, additional modeling and experiments will be required to validate this correlation hypothesis.

MATERIALS AND METHODS

Sample collection sites. Three of the seven major and permanent aquatic drainage systems of CCB (22) were sampled, the Churince Spring (Churince here), La Becerra/Garabatal River (Becerra here), and the Pozas Rojas system, taking water samples at several sites within each of these sites. A total of 11 water samples were analyzed: five from Churince, one from Becerra, and five from Pozas Rojas (Table 1). Each site was referenced by GPS to calculate the geographic distance between sampling locations. In addition, we analyzed the virus present in the intestinal contents of three different species of fish collected in Churince: the invasive cichlid fish from Africa, known as jewelfish (*Hemichromis guttatus*), and two native species, the mosquito fish, *Gambusia marshi*, from the *Poeciliidae*, native to northeastern Mexico, and the pupfish, *Cyprinodon bifasciatus*, endemic to the springs of CCB, from the *Cyprinodontidae* family.

Sample concentration, nucleic acid isolation, and sequencing. Each sample consisted of 7 to 11 liters of water collected in sterile polypropylene bottles. The concentration process was performed within 12 h after the time of collection by tangential ultrafiltration (50). Each sample was separately filtered through a polysulfone filter (F80A; Fresenius) with a 15- to 20-kDa cutoff point, previously washed with a solution of sodium pyrophosphate 0.1% using a peristaltic pump (Masterflex; Cole-Parmer) to recirculate water at a flow rate of 1,700 ml/min. Samples were concentrated by a factor of approximately 100× and stored at −20°C until use.

Prior to nucleic acid extraction, the concentrated water samples were subjected to a series of centrifugations. First, salt crystals and other debris were removed by low-speed centrifugation (1,800 × *g* for 30 min). The supernatants were subjected to repeated ultracentrifugation at 200,000 × *g* for 2 h in an SW40 rotor (Beckman), keeping the pellets. The final pellets were combined and resuspended in phosphate-buffered saline (PBS) and subjected to an additional centrifugation at 250,000 × *g* for 2 h in an SW55 rotor (Beckman). The supernatant was decanted and the pellet was resuspended in 450 μl of PBS by overnight incubation at 4°C. Before nucleic acid extraction, the resuspended samples were treated with Turbo DNase (Ambion, USA) and RNase (Sigma, USA) for 30 min at 37°C and immediately chilled on ice.

Nucleic acids were extracted from 200 μl of material using a PureLink viral RNA/DNA kit according to the manufacturer's instructions (Invitrogen, USA), employing linear acrylamide (AM9520; Ambion) instead of yeast tRNA as the carrier. Nucleic acids were eluted in nuclease-free water, aliquoted, quantified in a NanoDrop ND-1000 (NanoDrop Technologies, DE), and stored at −70°C until further use. To extract nucleic acids from the intestinal content of the investigated fish species, 200 μl of intestinal content was added to conical screw-cap tubes containing 100 mg of 150- to 212-μm glass beads (Sigma, USA), chloroform (100 μl), and PBS up to 1 ml. Samples were homogenized in a bead beater (Biospec Products, USA) and clarified by centrifugation at 2,000 × *g* for 10 min. The supernatants were recovered and filtered in Spin-X 22-μm-pore filters (Costar, NY, USA) at 5,000 × *g* for 10 to 20 min. The filtered samples were treated with Turbo DNase (Ambion, USA) and RNase (Sigma, USA), and the nucleic acid extraction was carried out as described above.

The random amplification of nucleic acids was performed essentially as described previously (51). Briefly, reverse transcription was done using SuperScript III reverse transcriptase (Invitrogen, USA) and primer-A (5′-GTTTCCCAGTAGGTCTCNNNNNNNNN-3′). The cDNA strand was generated by two rounds of synthesis with Sequenase 2.0 (USB, USA), followed by amplification with Phusion DNA polymerase (Finnzymes) using primer-B (5′-GTTTCCCAGTAGGTCTC-3′). As a negative control for spurious DNA amplification during processing of the samples, in an extra tube water was added in place of the sample in all sets of reactions carried out.

The amplified products were purified with a DNA Clean & Concentrator-5 kit (Zymo Research, USA) and digested with GsuI to remove 16 nucleotides from the 5′ and 3′ ends generated by PCR. The resulting DNA was purified once more by a DNA Clean & Concentrator-5 kit and used as input material for Illumina library construction. Briefly, purified DNA was fragmented to about 200 nucleotides using a Covaris M220 focused ultrasonicator, and libraries were prepared with a TruSeq DNA sample preparation kit (version LT) by following the manufacturer's protocol. Final libraries, including adaptors, were 318 to 400 nucleotides in length.

A total of 14 samples, five from Churince, one from Becerra, five from Pozas Rojas, and three from fish intestinal contents (FIC), were deep sequenced with an Illumina NextSeq500 sequencer, generating single reads of 125 bases in length. Fourteen libraries, one for each sample, were uniquely tagged, pooled, and then sequenced together. The image processing/base calling and demultiplexing of reads were performed by Illumina MiSeq-Control real-time analysis, version 1.18.54, and bcl2fastq, v2.15.0.4, respectively.

Analysis of the sequences and basic taxonomic and functional annotation. In-house scripts and programs were developed for the analysis of the reads, including the following steps. (i) PCR duplicate reads (up to 2% mismatches when aligned to its representative sequence) were removed from the data sets by using CD-HIT-DUP (-e 0.03 -m f), version 4.6.8. These reads are considered the result of sequencing two or more copies of the exact same DNA fragment. (ii) For each unique read, the adapter and the bases from the 5′ and 3′ ends with no-call sites (N residues) and with low-quality scores ($<$20, corresponding to an error rate of $>$1%) were trimmed. Low-complexity reads were then masked, and reads with more than 20 N nucleotides (either from sequencing error or as a result of applying a DUST algorithm) or less than 60 bp long were removed from the rest of the analysis. (iii) For bacterial rRNA removal, the program SMALT, version 0.7.5 (52), was used to align the reads against the bacterial rRNA database to remove them (parameters -n 8 -r 0 -y 0.9). The remaining sequences were considered valid reads. (iv) For taxonomic identification, first fast identification of candidate viral reads was done by aligning valid unique reads to the virus, and then bacterial and fungal nucleotide databases from NCBI (minimally nonredundant nucleotide) (53) were analyzed using SMALT (-n 8 -r 0 -y 0.7). The reads that mapped against these databases then were aligned with standalone BLASTn (-num_alignments 20 -evalue 0.001 -word_size 9) (54) against the nonredundant nucleotide database to remove reads that had higher identities to nonviral sequences (false-positive alignments). Reads that did not map using nucleotide alignment and that were greater than 120 bp in size were compared to the virus nonredundant protein database using BLASTx (-num_alignments 20 -evalue 0.00001) to look for novel viruses; viral hits were then compared to all nr databases to remove false-positive results. Finally, reads that did not map at the protein level with viruses were assembled using IDBA-UD (-mink 21 -maxk 75 -step 5 -min_contig 300) (55), and contigs greater than 300 bp were compared to all nr databases. Viral hits were then compared to all nr databases. (v) To assign reads to the most appropriate taxonomic level, considering that they might have multiple matches, the software MEGAN 5.10.6 (56) was used. Each read was assigned a node in the NCBI taxonomy according to the MEGAN 5 lowest common taxonomic ancestor (LCA) based on a set of its 20 best-scoring BLAST significant hits. This reduced the risk of false-positive matches, since species-specific sequences are assigned to taxa near the leaves of the NCBI tree, whereas widely conserved sequences are assigned to high-order taxa (genera, family, order, etc.). The parameters for the LCA algorithm were the following: min support, 2; min score, 50; top percent, 10. In addition, the two reads matching must be in two different regions of the same viral genome to be considered a positive viral hit. (vi) Taxonomic viral reads were functionally annotated using the SEED, level 1 system, and InterPro2GO, level 2 subsystem, of MEGAN 5. SEED reads are assigned hierarchies using mapping files that link RefSeq identifiers to identifiers in these functional classifications based on their best-alignment BLASTx hit. For InterPro2GO, proteins are mapped onto InterPro, and then these are placed in a GO-based tree.

Analysis of diversity and comparative metagenomics. To estimate viral diversity in each sample, contig spectra were generated using Circonspect (-v 120 -u 100 -s 10,000 -r 20) (6). A set of 10,000 random valid reads was extracted from each sample and assembled by the Minimo assembler using a minimum overlap of 35 bp and 98% sequence identity. Twenty repetitions were performed to generate an average contig spectrum. Sequences of less than 100 bp were discarded, and all other sequences were trimmed to 120 bp prior to assembly to obtain identical sequence size in the repeated assemblies. Settings not listed were not changed from defaults. Average contig spectra were used as inputs to the Phage Communities from Contig Spectra (PHACCS) tool (57). The viral diversity estimation was done by evaluating all rank abundance models and selected the best-fitting one. The average genome size used was 50 kbp in all cases.

Since geographical proximity is not necessarily a good predictor of how two viral communities might be related in CCB, in part due to the unknown hydrological connections at the deep aquifer level, viral taxonomic comparisons between samples were done. Statistical analyses were conducted in an R statistical environment (58) unless otherwise indicated, using multivariate community ecology procedures of the vegan package (59). First, nucleotide and protein taxonomic reads assigned to viral taxonomy were extracted from MEGAN 5 to generate count matrices. They were then normalized to the sample with the smallest number of total valid reads to represent relative abundances. Bray-Curtis dissimilarity distances were then computed from these matrices at species, genus, and family levels individually using the Canberra distance index (26). Canonical analysis of principal coordinates (PAC) was performed with the capscale function to test compositional differences between viral communities of each drainage system (60). The function Ordiellipse, based on standard deviations, was used to draw ellipses that illustrate the significant correlations found in the taxonomic profiles within each of the three aquatic systems in CCB (Churince, Becerra, and Pozas Rojas) and also in the FIC, using a confidence of 0.95. To confirm the differences in the viral composition between these systems, we then used a nonparametric multivariate permutation test (PERMANOVA) analysis (61) implemented with the Adonis R function, always using 999 permutations.

Metagenomic assembly and phylogeny. Metagenomic reads from specific viral families identified were *de novo* assembled using IDBA-UD (-mink 21 -maxk 75 -step 5 -min_contig 200) (55); consensus sequences were obtained and used to map all reads with the SMALT (-n 8 -r 0 -y 0.9) program to generate a better consensus sequence. Contigs homologous to gene markers that were at least 200 bp long then were phylogenetically characterized. The analysis required an approach different from that used to compare full-length proteins due to the fact that metagenomics sequences are fragmentary and not completely overlapping. Therefore, for each family, a database of complete proteins of interest was first created using all sequences available in GenBank as of January 2016. A representative subset at a 95% level of sequence identity derived from a CD-HIT analysis then was used (-c 95 -A 90) (62) in order to perform a reference alignment using the ClustalW method with the default pairwise alignment parameters. We next combined metagenomics contigs into a single large alignment by using the software MAFFT with the option align fragment sequences to reference alignment (mafft -addfragments fragments -reorder -thread -1 existing_alignment). Finally, maximum likelihood phylogenetic trees were generated with 500 repetition bootstraps using the MEGA v.6 program (TN93 with a proportional discrete Gamma distribution [+G], and a fraction of invariant sites [+I] was selected as the best-fitting model with the lowest Bayesian information criterion scores) for all cases.

Comparison of Cienegas viromes to other viral metagenomics data sets. The Churince and Pozas Rojas viromes were compared directly to available viromes on the Metavir website (http://metavir-meb.univ-bpclermont.fr/) by using the same taxonomic profiles, which are based on BLASTx results, as ours. These studies included freshwater (17, 45, 47, 63), marine water (6, 10, 46), and microbialites (35) and are summarized in Table 5. All profiles were imported into MEGAN 5 and then normalized to the sample with the smallest number of virus taxonomic reads. A dissimilarity species matrix then was computed from pairwise counts between CCB samples and these water samples using the Bray-Curtis distance. In order to take into account not only the known but also the unknown reads, species richness was also compared by using the number of different virotypes and the Shannon index values reported in those works. The same parameters of PHACCS used in the previous studies were used in our samples.

Accession number(s). HTS data are available in the NCBI database under accession number SRP136675. A MEGAN file for the virus-assigned reads derived from the HTS data for all samples is also available at the OneDrive site (https://1drv.ms/u/s!Ak9PFqzyLgkbgS42Z3kIMzDtLlxF).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.00465-18.

SUPPLEMENTAL FILE 1, PDF file, 0.4 MB.

ACKNOWLEDGMENTS

We thank the Instituto de Biotecnologia-UNAM for granting access to its computer cluster and Jerome Verleyen and Juan Manuel Hurtado for their computational support. We thank Laura Espinosa Asuar, Erika Aguirre Planter, and Marco Antonio Espinoza for technical assistance, the Protected Area APFF-Cuatro Ciénegas for their logistics and facilitation, Arturo Gonzales from Museo del Desierto de Saltillo for the permits to work in La Becerra, the people of Ejido del Venado for the permits to work in Pozas Rojas, the CBTA22 high school for allowing the use of their laboratory facilities, and the work of the high school students during the processing of the samples.

L.E.E. and V.S. worked on the manuscript during their sabbatical leave with a scholarship from PASPA-DGAPA UNAM.

This work was supported by grants WWF-FCS to V.S. and L.E.E.

REFERENCES

1. Nigro OD, Jungbluth SP, Lin H-T, Hsieh C-C, Miranda JA, Schvarcz CR, Rappé MS, Steward GF. 2017. Viruses in the oceanic basement. mBio 8:e02129-16. https://doi.org/10.1128/mBio.02129-16.
2. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp L, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S; Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti

E, Sullivan MB. 2015. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science 348:1261498. https://doi.org/10.1126/science.1261498.

3. Lu J, Chen F, Hodson RE. 2001. Distribution, isolation, host specificity, and diversity of cyanophages infecting marine Synechococcus spp. in river estuaries. Microbiology 67:3285–3290.
4. Clokie MRJ, Mann NH. 2006. Marine cyanophages and light. Environ Microbiol 8:2074–2082. https://doi.org/10.1111/j.1462-2920.2006.01171.x.
5. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99:14250–14255. https://doi.org/10.1073/pnas.202488399.
6. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F. 2006. The marine viromes of four oceanic regions. PLoS Biol 4:e368. https://doi.org/10.1371/journal.pbio.0040368.
7. Culley AI, Lang AS, Suttle CA. 2006. Metagenomic analysis of coastal RNA virus communities. Science 312:1795–1798. https://doi.org/10.1126/science.1127404.
8. Monier A, Claverie J-M, Ogata H. 2008. Taxonomic distribution of large DNA viruses in the sea. Genome Biol 9:R106. https://doi.org/10.1186/gb-2008-9-7-r106.
9. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M. 2011. Raw sewage harbors diverse viral populations. mBio 2:1–11. https://doi.org/10.1128/mBio.00180-11.
10. Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, McCrow JP, Tovchigrechko A, Yooseph S, Venter JC. 2012. Metagenomic exploration of viruses throughout the Indian Ocean. PLoS One 7:e42047. https://doi.org/10.1371/journal.pone.0042047.
11. Mizuno CM, Rodriguez-valera F, Kimes NE, Ghai R. 2013. Expanding the marine virosphere using metagenomics. PLoS Genet 9:e1003987. https://doi.org/10.1371/journal.pgen.1003987.
12. Winter C, Garcia JAL, Weinbauer MG, DuBow MS, Herndl GJ. 2014. Comparison of deep-water viromes from the Atlantic Ocean and the Mediterranean Sea. PLoS One 9:1–8. https://doi.org/10.1371/journal.pone.0100600.
13. Kim Y, Aw TG, Teal TK, Rose JB. 2015. Metagenomic investigation of viral immunities in ballast water. Environ Sci Technol 49:8396–8407. https://doi.org/10.1021/acs.est.5b01633.
14. Nunes-Alves C. 2015. Marine microbiology: deep sequencing of the global oceans. Nat Rev Microbiol 16:378. https://doi.org/10.1038/nrg3971.
15. Djikeng A, Kuzmickas R, Anderson NG, Spiro DJ. 2009. Metagenomic analysis of RNA viruses in a fresh water lake. PLoS One 4:e7264. https://doi.org/10.1371/journal.pone.0007264.
16. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M. 2009. Metagenomic analysis of viruses in reclaimed water. Environ Microbiol 11:2806–2820. https://doi.org/10.1111/j.1462-2920.2009.01964.x.
17. Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C. 2013. Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. ISME J 7:359–369. https://doi.org/10.1038/ismej.2012.101.
18. Mohiuddin M, Schellhorn HE. 2015. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. Front Microbiol 6:960. https://doi.org/10.3389/fmicb.2015.00960.
19. Wood-Charlson EM, Weynberg KD, Suttle CA, Roux S, Van Oppen MJH. 2015. Metagenomic characterization of viral communities in corals: mining biological signal from methodological noise. Environ Sci Technol 17:3440–3449.
20. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437. https://doi.org/10.1038/nature12352.
21. Stein BA, Kutner LS, Adams JS. 2000. Precious heritage: the status of biodiversity in the United States. Oxford University Press, Oxford, United Kingdom.
22. Minckley W. 1969. Environments of the Bolson of Cuatro Cienegas, Coahuila, Mexico. Sci Ser 2:1–65.
23. Cole GA. 1984. Crustacea from the Bolson of Cuatro Cienegas, Coahuila, Mexico. J Arizona Nevada Acad Sci Biota Cuatro Cienegas 19:3–12.
24. Lemos-Espinal JA, Smith GR. 2016. Amphibians and reptiles of the state of Coahuila, Mexico, with comparison with adjoining states. Zookeys 593:117–137. https://doi.org/10.3897/zookeys.593.8484.
25. Souza V, Espinosa-Asuar L, Escalante AE, Eguiarte LE, Farmer J, Forney L, Lloret L, Rodríguez-Martínez JM, Soberón X, Dirzo R, Elser JJ. 2006. An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. Proc Natl Acad Sci U S A 103:6565–6570. https://doi.org/10.1073/pnas.0601434103.
26. Escalante AE, Eguiarte LE, Espinosa-Asuar L, Forney LJ, Noguez AM, Souza Saldivar V. 2008. Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. FEMS Microbiol Ecol 65:50–60. https://doi.org/10.1111/j.1574-6941.2008.00496.x.
27. Cerritos R, Eguiarte LE, Avitia M, Siefert J, Travisano M, Rodríguez-Verdugo A, Souza V. 2011. Diversity of culturable thermo-resistant aquatic bacteria along an environmental gradient in Cuatro Ciénegas, Coahuila, México. Antonie Van Leeuwenhoek 99:303–318. https://doi.org/10.1007/s10482-010-9490-9.
28. Bonilla-Rosso G, Peimbert M, Alcaraz LD, Hernández I, Eguiarte LE, Olmedo-Alvarez G, Souza V. 2012. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin II: community structure and composition in oligotrophic environments. Astrobiology 12:659–673. https://doi.org/10.1089/ast.2011.0724.
29. Moreno-Letelier A, Olmedo-Alvarez G, Eguiarte LE, Souza V. 2012. Divergence and phylogeny of firmicutes from the Cuatro Ciénegas Basin, Mexico: a window to an ancient ocean. Astrobiology 12:674–684. https://doi.org/10.1089/ast.2011.0685.
30. López-Lozano NE, Eguiarte LE, Bonilla-Rosso G, García-Oliva F, Martínez-Piedragil C, Rooks C, Souza V. 2012. Bacterial communities and the nitrogen cycle in the gypsum soils of Cuatro Ciénegas Basin, Coahuila: a Mars analogue. Astrobiology 12:699–709. https://doi.org/10.1089/ast.2012.0840.
31. Tapia-Torres Y, Elser JJ, Souza V, García-Oliva F. 2015. Ecoenzymatic stoichiometry at the extremes: how microbes cope in an ultra-oligotrophic desert soil. Soil Biol Biochem 87:34–42. https://doi.org/10.1016/j.soilbio.2015.04.007.
32. Alcaraz LD, Olmedo G, Bonilla G, Cerritos R, Hernández G, Cruz A, Ramírez E, Putonti C, Jiménez B, Martínez E, López V, Arvizu JL, Ayala F, Razo F, Caballero J, Siefert J, Eguiarte L, Vielle J-P, Martínez O, Souza V, Herrera-Estrella A, Herrera-Estrella L. 2008. The genome of Bacillus coahuilensis reveals adaptations essential for survival in the relic of an ancient marine environment. Proc Natl Acad Sci U S A 105:5803–5808. https://doi.org/10.1073/pnas.0800981105.
33. Moreno-Letelier A, Olmedo G, Eguiarte LE, Martinez-Castilla L, Souza V. 2011. Parallel evolution and horizontal gene transfer of the pst operon in firmicutes from oligotrophic environments. Int J Evol Biol 2011:781642. https://doi.org/10.4061/2011/781642.
34. Peimbert M, Alcaraz LD, Bonilla-Rosso G, Olmedo-Alvarez G, García-Oliva F, Segovia L, Eguiarte LE, Souza V. 2012. Comparative metagenomics of two microbial mats at Cuatro Ciénegas Basin I: ancient lessons on how to cope with an environment under severe nutrient stress. Astrobiology 12:648–658. https://doi.org/10.1089/ast.2011.0694.
35. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F. 2008. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452:340–343. https://doi.org/10.1038/nature06735.
36. Souza V, Falcón LI, Elser JJ, Eguiarte LE. 2007. Protecting a window into the ancient Earth: towards a Precambrian park at Cuatro Cienegas, Mexico. Evol Ecol Res http://www.evolutionary-ecology.com/citizen/Towards%20a%20Precambrian%20Park%20at%20Cuatro%20Cienegas.pdf.
37. Rebollar EA, Avitia M, Eguiarte LE, González-González A, Mora L, Bonilla-Rosso G, Souza V. 2012. Water-sediment niche differentiation in ancient marine lineages of Exiguobacterium endemic to the Cuatro Cienegas Basin. Environ Microbiol 14:2323–2333. https://doi.org/10.1111/j.1462-2920.2012.02784.x.
38. Lee ZMP, Poret-Peterson AT, Siefert JL, Kaul D, Moustafa A, Allen AE, Dupont CL, Eguiarte LE, Souza V, Elser JJ. 2017. Nutrient stoichiometry shapes microbial community structure in an evaporitic shallow pond. Front Microbiol 8:949. https://doi.org/10.3389/fmicb.2017.00949.
39. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, Sabunciyan S, Talbot CC, Prandovszky E, Gurnon JR, Agarkova IV, Leister F, Gressitt KL, Chen O, Deuber B, Ma F, Pletnikov MV, Van Etten

JL. 2014. Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. Proc Natl Acad Sci U S A 111:16106–16111. https://doi.org/10.1073/pnas.1418895111.

40. Petro TM, Agarkova IV, Zhou Y, Yolken RH, Van Etten JL, Dunigan DD. 2015. Response of mammalian macrophages to challenge with the chlorovirus Acanthocystis turfacea chlorella virus 1. J Virol 89:12096–12107. https://doi.org/10.1128/JVI.01254-15.
41. Correa AMS, Ainsworth TD, Rosales SM, Thurber AR, Butler CR, Vega Thurber RL. 2016. Viral outbreak in corals associated with an in situ bleaching event: atypical herpes-like viruses and a new megavirus infecting Symbiodinium. Front Microbiol 7:127.
42. Halary S, Temmam S, Raoult D, Desnues C. 2016. Viral metagenomics: are we missing the giants? Curr Opin Microbiol 31:34–43. https://doi.org/10.1016/j.mib.2016.01.005.
43. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed). 2012. Virus taxonomy. Classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego, CA.
44. Hernández A, Espinosa-Pérez HS, Souza V. 2017. Trophic analysis of the fish community in the Ciénega Churince, Cuatro Ciénegas, Coahuila. PeerJ 5:e3637. https://doi.org/10.7717/peerj.3637.
45. Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P, Ruggles-Sage B, Romer Z, Putonti C. 2016. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. Mar Freshw Res 67:1700–1708. https://doi.org/10.1071/MF15172.
46. Hurwitz BL, Sullivan MB. 2013. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS One 8:e57355. https://doi.org/10.1371/journal.pone.0057355.
47. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One 7:e33641. https://doi.org/10.1371/journal.pone.0033641.
48. Thingstad TF. 2000. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. Limnol Oceanogr 45:1320–1328. https://doi.org/10.4319/lo.2000.45.6.1320.
49. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, Coutinho FH, Dinsdale EA, Felts B, Furby KA, George EE, Green KT, Gregoracci GB, Haas AF, Haggerty JM, Hester ER, Hisakawa N, Kelly LW, Lim YW, Little M, Luque A, McDole-Somera T, McNair K, de Oliveira LS, Quistad SD, Robinett NL, Sala E, Salamon P, Sanchez SE, Sandin S, Silva GGZ, Smith J, Sullivan C, Thompson C, Vermeij MJA, Youle M, Young C, Zgliczynski B, Brainard R, Edwards RA, Nulton J, Thompson F, Rohwer F. 2016. Lytic to temperate switching of viral communities. Nature 531:466–470. https://doi.org/10.1038/nature17193.
50. Polaczyk AL, Narayanan J, Cromeans TL, Hahn D, Roberts JM, Amburgey JE, Hill VR. 2008. Ultrafiltration-based techniques for rapid and simultaneous concentration of multiple microbe classes from 100-L tap water samples. J Microbiol Methods 73:92–99. https://doi.org/10.1016/j.mimet.2008.02.014.
51. Taboada B, Espinoza MA, Isa P, Aponte FE, Arias-Ortiz MA, Monge-Martínez J, Rodríguez-Vázquez R, Díaz-Hernández F, Zárate-Vidal F, Wong-Chew RM, Firo-Reyes V, Del Río-Almendárez CN, Gaitán-Meza J, Villaseñor-Sierra A, Martínez-Aguilar G, Salas-Mier MDC, Noyola DE, Pérez-Gónzalez LF, López S, Santos-Preciado JI, Arias CF. 2014. Is there still room for novel viral pathogens in pediatric respiratory tract infections? PLoS One 9:e113570. https://doi.org/10.1371/journal.pone.0113570.
52. Ponstingl H. 2012. SMALT–Wellcome Trust Sanger Institute, 0.7.6. Wellcome Trust Sanger Institute, Hinxton, United Kingdom.
53. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res 42:D553–D559. https://doi.org/10.1093/nar/gkt1274.
54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
55. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428. https://doi.org/10.1093/bioinformatics/bts174.
56. Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res 21:1552–1560. https://doi.org/10.1101/gr.120618.111.
57. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, Felts B, Nulton J, Mahaffy J, Rohwer F. 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6:41. https://doi.org/10.1186/1471-2105-6-41.
58. R Development Core Team. 2014. R: a language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna, Austria.
59. Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlin D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry M, Stevens H, Szoecs E, Wagner H. vegan: community ecology package. R package version. 1.17.11. https://CRAN.R-project.org/package=vegan.
60. Anderson MJ, Willis TJ. 2003. Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology 84:511–525. https://doi.org/10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2.
61. Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46. https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x.
62. Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
63. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. Science 326:858–861. https://doi.org/10.1126/science.1179287.
64. Held NL, Whitaker RJ. 2009. Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. Environ Microbiol 11:457–466. https://doi.org/10.1111/j.1462-2920.2008.01784.x.