UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE ECOLOGÍA

DIVERSIDAD Y EVOLUCIÓN DE VIRUS EN LA CUENCA DE CUATRO CIÉNEGAS: DOMOS DEL ARQUEANO

TESIS
QUE PARA OBTENER EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA
ALEJANDRO MIGUEL CISNEROS MARTÍNEZ

DIRECTOR DE TESIS
DRA. VALERIA SOUZA SALDÍVAR
INSTITUTO DE ECOLOGÍA
COMITÉ TUTOR
DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS
DR. LUIS DAVID ALCARAZ PERAZA
FACULTAD DE CIENCIAS

CIUDAD UNIVERSITARIA, CD. MX., NOVIEMBRE DE 2023

A Andrea, André y Penny

A mis padres, mi hermano y mi familia

A mis amigos y compañeros

A mis maestros y tutores

Agradecimientos

El presente trabajo se realizó bajo la dirección de la Dra. Valeria Souza Saldívar, en el Laboratorio de Evolución Molecular y Experimental del Departamento de Ecología Evolutiva del Instituto de Ecología de la Universidad Nacional Autónoma de México.

Se agradece la asesoría y seguimiento durante el desarrollo de la tesis por parte de los miembros del comité tutor conformado por la Dra. Valeria Souza Saldívar, el Dr. Arturo Carlos II Becerra Bracho, y el Dr. Luis David Alcaraz Peraza.

Se agradecen los comentarios y aportaciones por parte de los miembros del jurado conformado por el Dr. Carlos Federico Arias Ortiz, el Dr. Víctor Manuel González Zúñiga, la Dra. Eria Alaide Rebollar Caudillo, la Dra. Sonia Dávila Ramos, y el Dr. Arturo Carlos II Becerra Bracho.

Se agradece el apoyo técnico y teórico recibido por el Dr. Luis Enrique Eguiarte Fruns, la Dra. Rosalinda Tapia López, Ulises Erick Rodríguez Cruz, Manuel Rosas Barrera, Jazmín Sánchez Pérez, y David Madrigal Trejo.

Alejandro Miguel Cisneros Martínez fue alumno del Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), y fue beneficiado por el programa de becas nacionales de CONAHCYT (Número: 814975). La tesis forma parte de los proyectos de investigación financiados por el Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) con número IG200319 y IN204822.

Dedicatoria

Dedico esta tesis al amor de mi vida y la persona más importante para mí, mi esposa Andrea Teresa Téllez Galicia, a quien le tengo un amor incondicional y quien ha sido mi principal apoyo en esta aventura de vida que hemos decidido emprender juntos. A mi hijo André Alejandro Cisneros Téllez que, estoy seguro, nos ha dado el año más difícil de nuestras vidas, pero también el más lleno de satisfacciones. A mis padres, Miguel Cisneros Ramírez y Luz María Martínez Mejía, que no solo me han apoyado a lo largo de mi vida, sino que además han sido una inspiración y un modelo para seguir de responsabilidad, honestidad, amor y trabajo. A mi hermano, Héctor Miguel Cisneros Martínez, quien siempre me motiva para seguir por mi camino y alcanzar mi mayor realización. A mi nueva familia, Andrés Téllez Acosta, Celia Teresa Galicia García y Emilio Antonio Téllez Galicia, quienes me recibieron con los brazos abiertos durante la pandemia y siempre nos han demostrado un gran apoyo y cariño. A mis amigos de toda la vida, David Antonio Carnero Bustos y Jaime Elías Mochan Quesnel, que son como otros dos hermanos para mí. A mis amigos, Carlos Antonio González Palma, Armando Rodríguez Velasco y Oscar Said Quiroz Zerecero, cuya amistad me es muy importante. A mi tutora la Dra. Valeria Souza Saldívar, al Dr. Luis Eguiarte y a la Dra. Rosalina Tapia, quienes me han brindado un gran apoyo y me han demostrado una generosidad ilimitada. A todos mis amigos y compañeros, miembros del Laboratorio de Evolución Molecular y Experimental, incluyendo a Ulises Rodríguez, quienes se han expresado con un espíritu colaborativo y de compañerismo extraordinarios. Al Dr. Antonio Lazcano y al Dr. Arturo Becerra, con quienes trabajé desde el servicio social hasta el comienzo del doctorado, y quienes me han seguido apoyando con absoluta generosidad. Y, por último, a todos mis amigos y excompañeros del Laboratorio de Origen de la Vida, incluyendo al Dr. José Alberto Campillo, el Dr. Rodrigo Jácome, el Dr. Ricardo Hernández, a Abelardo Aguilar, Adrián Cruz y Coral Cruz, con quienes estoy muy agradecido por todos los años en los que compartimos seminarios y buenos momentos, y con quienes espero seguir colaborando en años por venir.

ÍNDICE

Resumen

La poza hipersalina Domos del Arqueano (AD), en la Cuenca de Cuatro Ciénegas (CCB), presenta concentraciones de sal que pueden variar desde el 5.3%, durante la temporada lluviosa, hasta concentraciones >35% durante la temporada seca, y es conocida por albergar una gran diversidad de arqueas que podría estar relacionada con la presencia de virus. Con el fin de describir la estructura y diversidad de la comunidad viral de AD a lo largo de las temporadas, y obtener indicios sobre el origen de la diversidad en este sitio, se tomaron seis muestras entre 2016 y 2019, alternando entre la temporada lluviosa y seca, así como seis muestras en 2020 a 0, 30 y 50 cm de profundidad. Esto resultó en 12 metagenomas que se compararon con 35 metagenomas públicos derivados de una variedad de ambientes con diferentes salinidades, y de otros sitios de CCB. La profundidad mostró una mayor influencia que las temporadas sobre la estructura y diversidad de una comunidad viral abundante en haloarqueavirus, que resultó ser la más diversa de los sitios comparados. También se encontró una mayor similitud con otros sitios de CCB y un incremento de la diversidad a mayores profundidades. Estos resultados apoyan la hipótesis de que la diversidad microbiana de CCB proviene del océano antiguo atrapado en el acuífero profundo hace cientos de millones de años. Con el fin de caracterizar los virus de AD, se ensamblaron virus a partir de los 12 metagenomas y se realizaron predicciones de hospederos. Primero se realizó una evaluación comparativa de ocho programas basados en cinco métodos diferentes sobre una base de datos de 1,046 pares conocidos de virus-hospedero. RaFAH, un programa que usa Random Forest y una base de datos de virus con hospedero conocido, presentó la mayor precisión (95.7%). Para predecir los hospederos de virus de AD se utilizó RaFAH, PHP (mejor herramienta basada en k-meros virus-hospedero), CrisprOpenDB (mejor herramienta basada en espaciadores CRISPR conocidos), y CrisprCustomDB (herramienta diseñada para usar espaciadores CRISPR predichos a partir de los mismos metagenomas de donde se ensamblaron los virus). Las predicciones sugieren que los virus de AD pueden infectar microorganismos halófilos, halotolerantes, alcalófilos, termófilos, oligotróficos, reductores de sulfato y marinos. CrisprCustomDB y PHP (base de datos local) mostraron la mayor concordancia entre sí, así como coherencia entre el ambiente y la taxonomía, hábitat, estilo de vida o metabolismo de los hospederos predichos, criterios que se pueden usar para mejorar la precisión de las predicciones, especialmente en comunidades virales muy diversas.

Abstract

The hypersaline pond Archaean Domes (AD), located in the Cuatro Ciénegas Basin (CCB), presents salinity concentrations that can vary from 5.3% during the rainy season, to concentrations >35% during the dry season, and is known to host a high diversity of archaea that could be related to the presence of viruses. To describe the structure and diversity of the AD viral community across seasons, and to gain insight into the origin of diversity at this site, six samples were collected between 2016 and 2019, alternating between the wet and dry seasons, and six samples were collected in 2020 at 0, 30, and 50 cm depth. This resulted in 12 metagenomes that were compared to 35 public metagenomes derived from a variety of environments with different salinities, and from other CCB sites. Depth showed a greater influence than season on the structure and diversity of a haloarchaeavirus-abundant viral community, which was found to be the most diverse of the sites compared. Greater similarity to other CCB sites and an increase in diversity with depth were also found. These results support the hypothesis that the microbial diversity of the CCB originates from the ancient ocean trapped in the deep aquifer hundreds of millions of years ago. To characterize the AD viruses, viruses were assembled from the 12 metagenomes and host predictions were made. First, a comparative evaluation of eight programs based on five different methods was performed on a database of 1,046 known virus-host pairs. RaFAH, a program using Random Forest and a database of viruses with known hosts, had the highest precision (95.7%). RaFAH, PHP (best tool based on virus-host k-mers), CrisprOpenDB (best tool based on known CRISPR spacers), and CrisprCustomDB (tool designed to use CRISPR spacers predicted from the same metagenomes from which the viruses were assembled) were used to predict AD virus hosts. Predictions suggest that AD viruses can infect halophilic, halotolerant, alkalophilic, thermophilic, oligotrophic, sulfate-reducing, and marine microorganisms. CrisprCustomDB and PHP (local database) showed the highest agreement with each other, as well as consistency between environment and predicted host taxonomy, habitat, lifestyle, or metabolism, criteria that can be used to improve the precision of predictions, especially in highly diverse viral communities.

Capitulo 1. Diversidad viral en los Domos del Arqueano

Introducción

Diversidad microbiana en ambientes hipersalinos

Los ambientes hipersalinos se encuentran ampliamente distribuidos en el mundo, principalmente en zonas áridas y semi-áridas, la mayoría en el hemisferio norte, representando aproximadamente el 44% del volumen de aguas continentales del planeta (Saccò et al., 2021). El Gran Lago Salado de Utah, el Salar de Atacama, el Mar Muerto y el Mar Caspio son algunos de los ambientes salados más conocidos del mundo. En México cuatro de los ocho lagos más grandes son salados. El Lago de Cuitzeo, Michoacán es el lago salado más grande de México. Otros lagos salados incluyen la Laguna de Alchichica en Puebla, Guerrero Negro en Baja California Sur, el Lago de Texcoco en el Estado de México y, la Laguna Grande y la Laguna Salada en la Cuenca de Cuatro Ciénegas (Alcocer & Hammer, 1998).

Los ambientes hipersalinos se caracterizan por concentraciones de sal superiores a las del agua de mar (3-4%) (DasSarma & DasSarma, 2012) y suelen clasificarse según sus niveles de salinidad; desde baja hipersalinidad (<10% NaCl), pasando por hipersalinidad intermedia (10%-20% NaCl), hasta alta hipersalinidad (>20% NaCl) (Ventosa et al., 2014). A su vez, los organismos halófilos se clasifican según la concentración de NaCl que requieren para un crecimiento óptimo: i) halófilos leves (1-5%); ii) halófilos moderados (5-20%), y iii) halófilos extremos (20-30%) (DasSarma & DasSarma, 2012). Algunos microorganismos halófilos también son organismos alcalófilos (Litchfield, 2011). Los alcalófilos pueden prosperar en ambientes con pH > 9 (Merino, 2019), sin embargo, los organismos haloalcalófilos se encuentran típicamente en lagos salinos alcalinos, también conocidos como lagos de sosa. Los lagos de sosa se caracterizan por aguas altamente alcalinas (pH > 9) como resultado de una alta alcalinidad de carbonatos (altas concentraciones de CO_3^{2-} y HCO_3^-) junto con bajas concentraciones de Ca^{2+} y Mg^{2+} (Kempe & Kazmierczak, 2011; Boros & Kolpakova, 2018).

De forma similar a los ambientes talásicos (con asociación marina) hipersalinos con pH neutro, la composición de la comunidad microbiana de los lagos de sosa está fuertemente influenciada por la salinidad, donde las mayores concentraciones de sal dan lugar a una gran abundancia de arqueas halófilas extremas pertenecientes a la clase *Halobacteria* y, por tanto, a una comunidad menos diversa (McGenity & Oren, 2012; Vavourakis et al., 2016; Castelán-Sánchez et al., 2019). Sin embargo, los lagos de sosa tienden a ser más diversos que los ambientes hipersalinos de pH neutro, lo que probablemente esté relacionado con la alta disponibilidad de CO_2 para los productores primarios (Jones et al., 1998) y las bajas concentraciones de Ca^{2+} y Mg^{2+} (Vavourakis et al., 2016). Los lagos de sosa están poblados por diversas bacterias y arqueas adaptadas a la salinidad y al pH, entre las que se incluyen miembros del grupo *Halomonas*, cepas bien representadas relacionadas con *Bacillus alcalophilus* (Jones et al., 1998) y, a mayores concentraciones de sal, *Euryarchaeota* de la clase *Halobacteria*, del orden *Methanosarcinales* (Vavourakis et al., 2016) y de los géneros *Natronococcus* y *Natronobacterium* (Jones et al., 1998; Litchfield, 2011).

Virus en ambientes hipersalinos

Los virus, que son conocidos como las entidades más abundantes y diversas del mundo, superando típicamente la abundancia bacteriana 10 veces, incluso en ambientes oligotróficos (Wommack & Colwell, 2000; Sullivan et al., 2017), también son abundantes en ambientes hipersalinos ($4x10^8$ - $2x10^9$ VLP/mL) (Santos et al., 2007). Los virus son actores clave en las comunidades microbianas, donde pueden aumentar la diversidad genotípica mediando el intercambio genético entre cepas bacterianas vía transducción, o también matando selectivamente a la población más densa, abundante y de rápido crecimiento (i. e. "kill the winner" y escenarios evolutivos similares) (Wommack & Colwell, 2000; Winter et al., 2010).

Se considera que la regulación de las poblaciones de bacterioplancton desde niveles tróficos superiores (*top-down control*, en inglés) es mediada principalmente por el pastoreo de protistas –también llamado bacterivoría– o por la lisis viral. Por un lado, los protistas heterotróficos ayudan a transportar la materia orgánica de los productores primarios hacia niveles tróficos superiores (Rocke et al., 2015), mientras que, por otro lado, la lisis viral es un mecanismo eficiente para el incremento del flujo de biomasa hacia la materia orgánica disuelta (la cual es fácilmente consumida por bacterias), disminuyendo así la transferencia de biomasa a niveles tróficos superiores (proceso conocido como derivación viral o *viral shunt*) (Wommack & Colwell, 2000; Sullivan et al., 2017). Si bien la contribución de la depredación por protistas como la de la lisis viral sobre la mortalidad bacteriana depende de las condiciones físico-químicas de cada sitio y del estado fisiológico de las poblaciones estudiadas (Medina et al., 2017), es posible que la lisis viral tenga un mayor impacto en condiciones oligotróficas, donde la biomasa consiste principalmente en procariontes. En el caso de ambientes hipersalinos, se ha observado que el pastoreo de protistas desaparece a salinidades superiores al 20% (Guixa-Boixareu et al., 1998), por lo que es de esperar que el control de las comunidades dominadas por arqueas halófilas sea principalmente mediado por virus de arqueas halófilas o haloarqueavirus.

Las tendencias de diversidad observadas en comunidades de microorganismos halófilos suelen extenderse a los virus que los infectan. Por ejemplo, el aumento en la abundancia de arqueas halófilas a lo largo de un gradiente creciente de salinidad suele ir acompañado por un incremento en la abundancia de haloarqueavirus, lo cual resulta en una disminución de la diversidad en la comunidad (Roux et al., 2016). Los haloarqueavirus pertenecen principalmente al antes orden *Caudovirales* (antes familias *Siphoviridae*, *Myoviridae* y *Podoviridae*) (abolición propuesta y aprobada en 2021, y ratificada en 2022 por el comité ejecutivo del Comité Internacional de Taxonomía de Virus (ICTV, por sus siglas en inglés) (Turner et al., 2023)) con una proporción menor perteneciente a otras familias virales como *Sphaerolipoviridae*, *Pleolipoviridae* y *Fuselloviridae* (Prangishvili et al., 2017).

La Cuenca de Cuatro Ciénegas y los Domos del Arqueano

La Cuenca de Cuatro Ciénegas (CCB, por sus siglas en inglés) es un oasis amenazado en el desierto de chihuahua, México. Este sitio se caracteriza por proporciones N:P que van desde muy bajo fósforo (157:1) a muy bajo nitrógeno (2:1) (Souza et al., 2012). A pesar de su estatus oligotrófico, CCB es conocido por su gran biodiversidad, incluyendo animales, plantas, hongos y procariontes, y se ha sugerido que es un análogo de la Tierra primitiva por diferentes razones, incluyendo la antigüedad de sus sedimentos (Souza et al., 2006) y la abundancia local de estructuras organosedimentarias, como tapetes microbianos y estromatolitos, que son conspicuas en el registro fósil del eón Arqueano (Walsh, 2010). La noción de CCB como un modelo de la Tierra primitiva se ve reforzada por la presencia de microorganismos endémicos, adaptados a una estequiometría que recuerda al supereón Precámbrico tardío (Alcaraz et al., 2008), relacionados con organismos marinos de los que se estima que divergieron hace 770-680 y 202-160 millones de años (Moreno-Letelier et al., 2012; Souza et al., 2018). Además, estudios isotópicos han demostrado que los sistemas acuáticos de CCB están compuestos en gran parte por agua subterránea influenciada por una bolsa magmática la cual permite al agua profunda subir y formar las pozas (Wolaver et al., 2013). Todos estos datos sugieren que la diversidad de CCB ha evolucionado como resultado de una larga estabilidad ambiental de un acuífero profundo que recrea las condiciones de un océano antiguo (Wolaver et al., 2013; Souza et al., 2018). Si bien, los sistemas acuáticos de CCB pueden considerarse como análogos de un océano antiguo, y por lo tanto modelos para el estudio de procesos ecológico y evolutivos que tuvieron lugar hace millones de años, sus suelos áridos y salinos, ricos en yeso, pueden verse como análogos de ambientes marcianos que posiblemente albergaron vida en algún momento de su historia geológica (como el cráter Gale que es rico en yeso –sulfatos hidratados– y que presenta varios indicios de ser un lago seco) y, por lo tanto, su estudio desde un punto de vista Astrobiológico, puede contribuir en el descubrimiento de biofirmas asociadas a minerales que resultan de actividad acuática, los cuales pueden usarse en la búsqueda de vida en otros planetas (López-Lozano et al., 2012; Souza et al., 2012).

En la primavera de 2016, se descubrió en CCB una poza pequeña (50 x 25 m) y poco profunda caracterizada por pH y salinidad elevadas (Medina-Chávez et al., 2023). Este sitio en particular tiene cisternas elipsoidales llenas de agua anaranjada, denominadas "círculos anaranjados" (OC, por sus siglas en inglés), que permanecen húmedas todo el año y que son particularmente ricas en arcilla y limo. Alrededor de dichos OC, en la interfase con sedimentos más arenosos, se forman tapetes microbianos bajo una costra salina. El mismo año de su descubrimiento, luego de que una fuerte lluvia disolviera la costra salina, los tapetes microbianos empezaron a inflarse con gases anóxicos reminiscentes del eón Arqueano, como metano y sulfuro de hidrógeno (Medina-Chávez et al., 2023), formando estructuras en forma de domo (Figura 1). Debido a los tapetes microbianos elásticos y a que la atmósfera en el interior de los domos es similar a la del eón Arqueano, el yacimiento es conocido como "Domos del Arqueano" (AD, por sus siglas en inglés) (Medina-Chávez et al., 2023;

Espinosa-Asuar et al., 2022). Dado que hay bajas cantidades de metaloenzimas de cobre, lo que concuerda con la baja biodisponibilidad de cobre durante el eón Arqueano (Madrigal-Trejo, 2022), y que los ambientes de pH y salinidad extremos se consideran modelos de antiguos ecosistemas marcianos (Banciu & Sorokin, 2013; Minegishi, 2013), AD podría considerarse un posible modelo de comunidades terrestres muy antiguas y un sitio de interés astrobiológico.

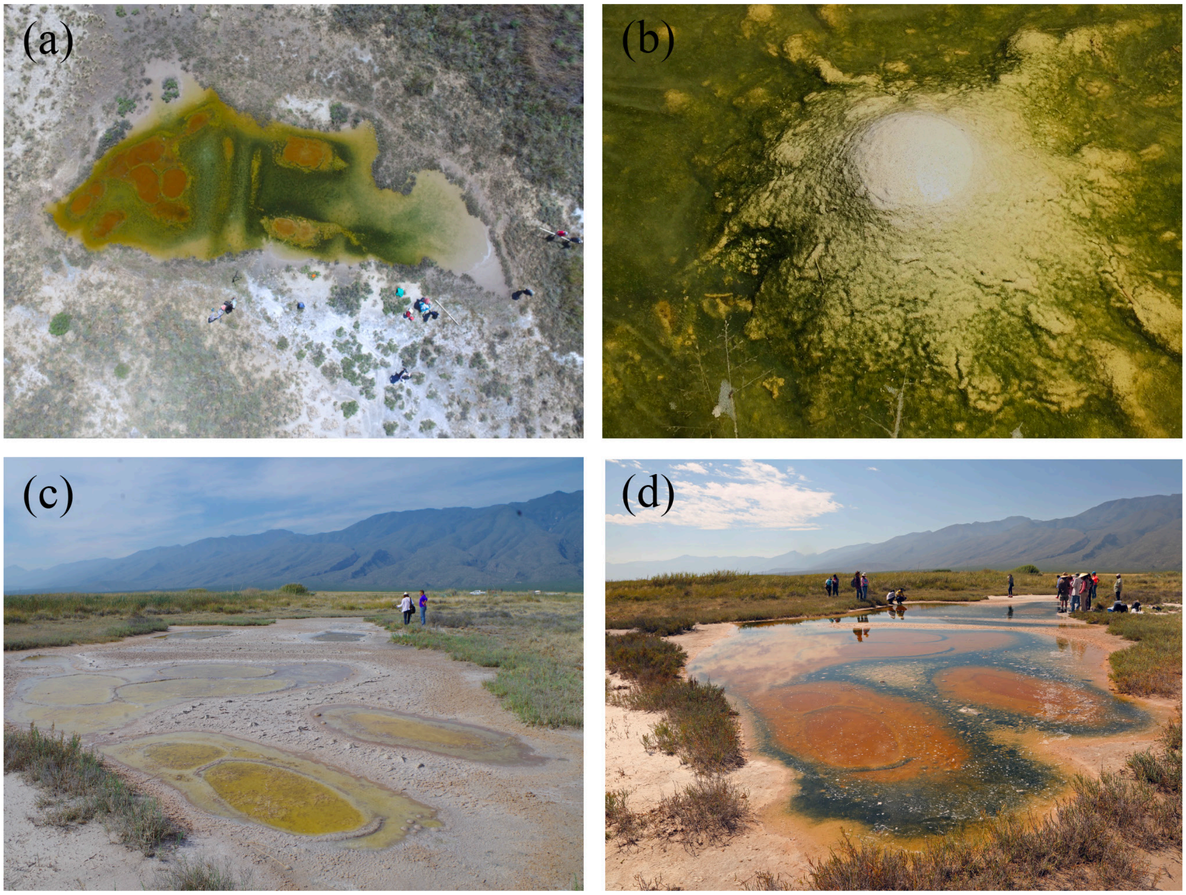


Figura 1. Poza Domos del Arqueano o AD en la Cuenca de Cuatro Ciénegas (CCB), México. (a) Vista aérea de la poza (50×25 m). (b) Tapete microbiano elástico formando una estructura en forma de domo con una costra salina en la parte superior. (c) AD durante la temporada seca (abril 2016) (d) AD durante la temporada lluviosa (septiembre 2019). En C y D son observan los círculos anaranjados u OC.

Antecedentes

Diversidad microbiana en AD

AD es un sitio que sufre grandes fluctuaciones estacionales que producen grandes variaciones entre la temporada húmeda y seca en la relación N:P (de 10:1 a 78:1), el pH (de 9.5 a 5.5) y la salinidad (de 5.3% a saturación) (Medina-Chávez et al., 2023; Espinosa-Asuar et al., 2022). A pesar de su naturaleza fluctuante y extrema, AD parece albergar una comunidad microbiana estacionalmente estable (Medina-Chávez et al., 2023), con cierto grado de taxones funcionalmente redundantes y una gran abundancia de genes de resistencia al pH y salinidad (Madrigal-Trejo et al., 2023). AD también alberga una gran diversidad de microorganismos, incluyendo más de 6000 variantes de secuencias de amplicón (ASV, por sus siglas en inglés) en 10 muestras obtenidas a una escala de 1.5 m (Espinosa-Asuar et al., 2022). Esta diversidad incluye una gran abundancia de bacterias halotolerantes, así como arqueas halófilas y metanogénicas (Medina-Chávez et al., 2023; Espinosa-Asuar et al., 2022). AD también tiene una gran riqueza y diversidad de arqueas, que son raras en el resto de CCB (Medina-Chávez et al., 2023; Espinosa-Asuar et al., 2022). Por último, una abundancia considerable de virus (~30% de las secuencias) (Medina-Chávez et al., 2023), sugiere que estos juegan un papel importante en el mantenimiento de la diversidad de arqueas en este sitio. Sin embargo, la estructura y diversidad de la fracción viral aun no ha sido explorada.

Virus en otros sitios de CCB

Diferentes estudios metagenómicos en CCB han descrito la comunidad viral dentro de estromatolitos del río Mezquites, trombolitos de la laguna Pozas Azules II (Desnues et al., 2008) y muestras de agua de las lagunas Churince, La Becerra y Pozas Rojas (Taboada et al., 2018). Estas comunidades están típicamente dominadas por bacteriófagos dsDNA de las familias *Siphoviridae*, *Myoviridae* y *Podoviridae*, pertenecientes al orden *Caudovirales,* seguidos por bacteriófagos ssDNA de la familia *Microviridae*, una variedad de virus DNA y RNA de diferentes familias virales eucariotas y virtualmente ningún virus que infecte arqueas (Taboada et al., 2018). En conjunto, estos estudios muestran que las comunidades virales de CCB tienden a reflejar los patrones de diversidad de sus hospederos, mostrando una alta diversidad dentro y entre sitios, similitud taxonómica con muestras marinas y fuertes señales de endemismo (Desnues et al., 2008; Taboada et al., 2018).

Justificación

Los virus son conspicuos en la mayoría de los ecosistemas del mundo y pueden desempeñar roles protagónicos en control de las poblaciones microbianas. La CCB ha demostrado ser un sitio único en el mundo que alberga una extraordinaria diversidad microbiana así como comunidades virales que reflejan los patrones de diversidad de sus hospederos bacterianos que son altamente diversos, presentan afinidades marinas y muestran fuertes señales de endemismo. La comunidad microbiana de AD ha mostrado una diversidad aún más extraordinaria que incluye una variedad de arqueas como no se había visto en ningún otro sitio de CCB. Estudiar los virus de AD puede brindarnos más pistas sobre el origen y mantenimiento de la diversidad microbiana en CCB y, más específicamente, en AD. Por otro lado, debido a que AD es un sitio hipersalino que muestra diferentes niveles de salinidiad a lo largo de las temporadas y a que la salinidad es uno de los parámetros que más influyen sobre el control de las poblaciones de microorganismos halófilos, es importante, no solo caracterizar sino también, contrastar los patrones de diversidad viral en AD con los de otros sitios del mundo con diferentes niveles de salinidad (incluyendo otros sitios de CCB) para tener un marco de referencia que nos permita evaluar la magnitud de las particularidades de este sitio.

Objetivos

General

Describir la estructura y diversidad de la comunidad viral de AD a lo largo de las temporadas, y obtener indicios sobre el origen de la diversidad en este sitio.

Particulares

1. Estimar la abundancia relativa de bacterias, arqueas y virus en cada una de las muestras de AD.

2. Describir la abundancia relativa y estimar la diversidad α de los virus de AD a diferentes niveles taxonómicos.
3. Evaluar la influencia de las temporadas lluviosa y seca sobre la estructura y diversidad de la comunidad viral de AD.
4. Valorar el efecto de la profundidad sobre la estructura y diversidad de la comunidad viral de AD.
5. Comparar la estructura y diversidad de la comunidad viral de AD con las de otros sitios con diferentes salinidades y de otras partes de CCB.

Hipótesis

Por un lado, debido a que AD es un sitio hipersalino que puede alcanzar el punto de saturación durante la temporada seca (Medina-Chávez et al., 2023), se esperaba encontrar una comunidad viral similar a la de otros sitios hipersalinos, en donde la abundancia de haloarqueavirus incrementa conforme aumenta la salinidad (Roux et al., 2016). A demás, debido a que la salinidad es uno de los parámetros que más afectan la estructura de estas comunidades microbianas (Roux et al., 2016), se esperaba que la estructura y diversidad de la comunidad viral se viera fuertemente influenciada por las grandes fluctuaciones en salinidad observadas entre las temporadas de lluvias y secas. Por otro lado, dado que AD presenta una comunidad microbiana altamente diversa y estable (Medina-Chávez et al., 2023), y que –por su ubicación– AD comparte una historia geológica con otros sitios de CCB, podía esperarse una comunidad viral altamente diversa con una estructura y diversidad independientes de las fluctuaciones ambientales, así como una comunidad viral con una estructura similar a las comunidades virales de otros sitios de CCB (Desnues et al., 2008; Taboada et al., 2018). A demás, debido a que el acuífero profundo parece ser una de las principales fuentes de agua de los sistemas acuáticos de CCB (Wolaver et al., 2013), así como de la gran diversidad de microorganismos con afinidad marina (Souza et al., 2006), se esperaba que las muestras colectadas a diferentes profundidades proporcionaran indicios sobre el origen de la diversidad en AD a partir del transporte de microorganismos desde el acuífero profundo, en consonancia con el modelo del “mundo perdido” –aislamiento ancestral y diversificación– sobre el origen de la diversidad en CCB (Souza et al., 2018).

Metodología

Toma de muestras

Las muestras se colectaron dentro del Rancho Pozas Azules (26°49'41,9 "N 102°01'23,6 "O) perteneciente a la Pronatura Noreste, en la Cuenca de Cuatro Ciénegas (CCB), en Coahuila, México, bajo el permiso científico de la SEMARNAT número SGPA/DGVS/03121/15.

Las muestras se recolectaron en abril de 2016, octubre de 2016, febrero de 2017, octubre de 2018, marzo de 2019, septiembre de 2019 y octubre de 2020. Las muestras tomadas entre febrero y abril corresponden a la estación seca, mientras que las tomadas entre septiembre y octubre corresponden a la estación húmeda. Para los tapetes microbianos, las muestras superficiales se recogieron mediante disección con bisturí estéril (8 cm^2 / 40 cm^3) y se transfirieron a tubos cónicos de 50 mL. Para las muestras más profundas, se utilizaron tubos de plástico de 30 cm como muestreadores de sedimentos a profundidades de 30 y 50 cm. Se recogieron tres muestras en los círculos anaranjados (OC): una muestra de agua superficial en un tubo cónico de 50 mL y dos más a profundidades de 30 y 50 cm, como se ha descrito anteriormente.

En total se tomaron 12 muestras (Tabla 1): tres muestras superficiales del tapete microbiano durante las temporadas secas (M1, M3, M5); cuatro muestras superficiales del tapete microbiano durante las temporadas húmedas (M2, M4, M6, D0); una muestra superficial de agua en el OC durante una temporada húmeda (C0); dos muestras profundas del tapete microbiano durante una temporada húmeda (D30, D50) y dos muestras profundas del OC durante una temporada húmeda (C30, C50). Todas las muestras se almacenaron en nitrógeno líquido hasta su procesamiento.

Tabla 1. Características de las muestras colectadas en los Domos del Arqueano, Cuatro Ciénegas, México.

Muestra	Procedencia	Profundidad (cm)	Mes	Año	Lecturas crudas	Lecturas filtradas
M1	Tapete microbiano	0	Abril	2016	28,859,454	26,799,269

M2	Tapete microbiano	0	Octubre	2016	4,772,053	4,412,620
M3	Tapete microbiano	0	Febrero	2017	8,203,484	7,484,431
M4	Tapete microbiano	0	Octubre	2018	10,030,782	9,442,166
M5	Tapete microbiano	0	Marzo	2019	25,873,990	24,402,939
M6	Tapete microbiano	0	Septiembre	2019	20,153,088	19,486,258
D0	Tapete microbiano	0	Octubre	2020	17,148,993	15,895,120
D30	Tapete microbiano	30	Octubre	2020	18,976,795	18,350,997
D50	Tapete microbiano	50	Octubre	2020	16,106,607	15,418,160
C0	Círculos anaranjados	0	Octubre	2020	24,065,589	22,124,414
C30	Círculos anaranjados	30	Octubre	2020	14,315,374	13,901,353
C50	Círculos anaranjados	50	Octubre	2020	18,050,094	17,604,941

Extracción de DNA y secuenciación

El DNA se extrajo de acuerdo con Purdy (2005) en el Laboratorio de Evolución Molecular y Experimental del Instituto de Ecología de la Universidad Nacional Autónoma de México, en la Ciudad de México. Brevemente, las extracciones siguieron un protocolo basado en columnas con un "Fast DNA Spin Kit for Soil" (MP Biomedical). El DNA total se envió al CINVESTAV-LANGEBIO, Irapuato, México, para la secuenciación *shotgun* con la tecnología Illumina Mi-Seq paired-end 2x300. El número de lecturas crudas y el número de lecturas después del filtrado de calidad se pueden ver en la tabla 1.

Descarga de datos metagenómicos

Con base en los metagenomas comparados en Roux et al., (2016), se descargaron del Sequence Read Archive (SRA) 35 metagenomas descritos en la literatura, y se agruparon en siete tipos de acuerdo con su procedencia (agua de Cuatro Ciénegas, microbialitas, hipersalino alto, hipersalino medio, hipersalino bajo, marinos y agua dulce). Los metagenomas incluyen 11 viromas de CCB aislados de las pozas Churince, La Becerra y Pozas Rojas (Taboada et al., 2018); tres viromas de microbialitas (Pozas Azules II y Río Mezquites de CCB, y Highborne cay, Bahamas) (Desnues et al., 2008); ocho viromas altamente hipersalinos derivados del lago Tyrell, Australia (Emerson et al., 2012), un viroma hipersalino de salinidad intermedia y otro de baja salinidad de la Bahía de San Diego, USA (Dinsdale et al., 2008a); seis viromas oceánicos (cuatro del océano pacífico y dos del océano atlántico) (Angly et al., 2006; Dinsdale et al., 2008b; McDaniel et al., 2008); y cinco viromas de agua dulce (dos lagos de Francia y tres estanques de una granja de tilapias en USA) (Dinsdale et al., 2008a; Roux et al., 2012). Los códigos de acceso del SRA correspondientes y el número de lecturas de cada metagenoma se pueden consultar en la tabla suplementaria *File S1* del Anexo 1.

Comparaciones taxonómicas de metagenomas

Las comparaciones de metagenomas se realizaron con un script personalizado (COMETS: COmpare METagenomeS. disponible en GitHub https://github.com/AleCisMar/COMETS). En resumen, este script toma como entrada una tabla de metadatos y todos los archivos FASTQ comprimidos con lecturas crudas *single end* o *paired end*. A continuación, utiliza fastp (Chen et al., 2018) para la eliminación de adaptadores, el filtrado de lecturas de baja calidad (se califican las lecturas con un máximo del 40% de bases con calidad < Q15) y la deduplicación de lecturas con todas las bases idénticas. A continuación, utiliza Kaiju (Menzel et al., 2016) para realizar clasificaciones taxonómicas comparando cada lectura de secuenciación con la base de datos nr_euk. Kaiju utiliza la transformada Burrows-Wheeler para buscar coincidencias entre secuencias de lecturas traducidas y la base de datos de genes codificantes microbianos, asignando el identificador taxonómico (de la taxonomía NCBI) con la coincidencia exacta más larga a cada lectura de secuenciación. Kaiju asigna el identificador taxonómico del ancestro menos común (LCA) si se encuentran coincidencias de la misma longitud en múltiples taxones (Menzel et al., 2016). Tras la clasificación taxonómica, COMETS produce una tabla de recuento y una tabla de taxonomía que se utilizan junto con la tabla de metadatos para construir un objeto phyloseq (McMurdie & Holmes, 2013) en R (R Core Team, 2021). En el siguiente paso produce curvas de rarefacción con el paquete myrlin (Cameron et al., 2021) y realiza una normalización por mediana de profundidad de secuenciación.

Por último, COMETS genera tres tipos de gráficos con ggplot2 (Wickham, 2016) a diferentes niveles taxonómicos: todos los identificadores taxonómicos asignados por Kaiju se consideran OTUs, los cuales pueden filtrarse según el nivel taxonómico, desde el filo hasta la especie. El primer tipo de gráfico es una barra apilada para la abundancia relativa de OTUs que representan al menos el 1% de las lecturas en al menos una muestra; el segundo tipo es un gráfico de puntos para la diversidad alfa (Shannon) y; el último es un gráfico de dispersión de escalado multidimensional no métrico (NMDS) para representar la diversidad beta (disimilitud de Bray-Curtis). Cuando se realizaron filtros taxonómicos, los OTUs sin la asignación taxonómica correspondiente (NA) se excluyeron de los cálculos de diversidad y de la generación de gráficos.

Para explorar las similitudes entre las comunidades virales en AD y otras comunidades virales en CCB y el resto del mundo, se calcularon las medidas de disimilitud de Bray-Curtis a partir de la tabla de recuento normalizada a diferentes niveles taxonómicos para construir árboles UPGMA con el programa NEIGHBOR del paquete PHYLYP, utilizando el orden de entrada y sin subreplicas (Felsenstein, 1989).

Análisis estadísticos de los perfiles taxonómicos

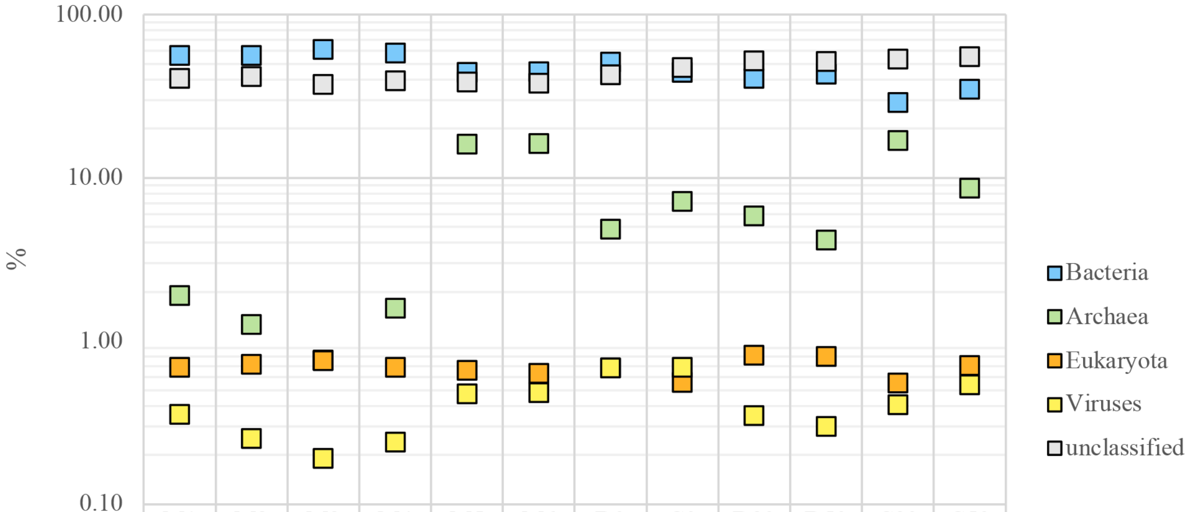
Para los análisis estadísticos iniciales, asumimos una distribución normal y realizamos ANOVA y pruebas t-Student de una cola con alfa = 0.05 para contrastar diferentes grupos de muestras: i) estaciones húmedas (M2, M4 y M6) y secas (M1, M3 y M5); ii) muestras superficiales antes de 2019 (M1, M2, M3 y M4), superficiales de 2019 en adelante (M5, M6, D0 y C0) y profundas (D30, D50, C30 y C50) y; iii) muestras superficiales (M1, M2, M3, M4, D0 y C0) y profundas (M5, M6, D30, D50, C30 y C50).

Para un análisis más robusto de la abundancia diferencial entre grupos de muestras, seguimos el pipeline IDEAmex desarrollado originalmente para el análisis de expresión diferencial (Jiménez-Jacinto et al., 2019) con el paquete edgeR (Robinson et al., 2010). La normalización se realizó con el método de la media recortada de los valores M o TMM, y los OTUs con abundancia diferencial se consideraron con un *log fold change* = 1.5 y un valor p = 0.01.

Resultados

Descripción de la comunidad microbiana en AD

Se analizó la abundancia relativa de lecturas asignadas a los tres dominios de la vida y a los virus para explorar la estructura taxonómica general de cada comunidad muestreada en AD. En promedio, la mayoría de las lecturas se asignaron a bacterias (46.95%), seguidas de lecturas sin clasificar (44.84%), arqueas (7.11%), eucariotas (0.69%) y virus (0.41%). Sin embargo, estas abundancias varían mucho entre las distintas muestras. Por ejemplo, las Bacterias oscilaron entre el 61.03% en la muestra superficial de la estación seca M3, y el 28.84% en la muestra de OC a 30 cm de profundidad C30, mientras que las Archaea variaron entre el 0.76% y el 16.88% en las mismas muestras (Figura 2).



	M1	M2	M3	M4	M5	M6	D0	C0	D30	D50	C30	C50
Bacteria	56.36	56.09	61.03	58.26	44.12	44.71	50.98	44.34	40.78	43.19	28.84	34.74
Archaea	1.89	1.26	0.76	1.58	16.11	16.21	4.85	7.16	5.83	4.16	16.88	8.60
Eukaryota	0.69	0.72	0.76	0.69	0.66	0.63	0.68	0.56	0.82	0.80	0.55	0.71
Viruses	0.35	0.25	0.19	0.24	0.47	0.48	0.68	0.69	0.35	0.30	0.40	0.54
unclassified	40.70	41.67	37.26	39.23	38.64	37.97	42.81	47.25	52.22	51.54	53.33	55.42

Figura 2. Porcentaje de lecturas asignadas a Bacteria, Arquea, Virus, o sin clasificar. Las muestras superficiales de 2016 a 2018 (M1-M4) tienen una mayor abundancia de lecturas asignadas a Bacteria y una menor abundancia de lecturas asignadas a Arquea y Virus. Los metagenomas muestreados a 30 y 50 cm de profundidad (D30, D50, C30, C50) muestran la mayor proporción de lecturas sin clasificar.

Para evaluar a qué factores pueden atribuirse estas diferencias entre muestras, comparamos las variaciones de abundancia relativa entre muestras agrupadas por estación o profundidad. No pudimos encontrar diferencias significativas entre las muestras de las temporadas seca (M1, M3 y M5) y húmeda (M2, M4 y M6) (Tabla 2). Sin embargo, hubo cambios significativos entre algunos grupos de muestras. Por ejemplo, las muestras de superficie de 2019 y 2020 (M5, M6, D0 y C0) se distinguen de las muestras de superficie de 2016 a 2018 (M1, M2, M3 y M4) por una menor proporción de lecturas asignadas a Bacterias (t-Student, t = 5.93, p = 0.0041), junto con una mayor abundancia relativa de lecturas asignadas a Archaea (p = 0.0312) y Virus (p = 0.0098). Las muestras de superficie a partir de 2019 sólo difieren significativamente de las muestras de profundidad (D30, D50, C30 y C50) por un menor porcentaje de lecturas no clasificadas (p = 0.0077).

Tabla 2. Media por grupos de muestras y los valores estadísticos de sus comparaciones. Valores p significativos (<0.05) se muestran en negritas.

	Dry	**Wet**	**t**	**p**	**Before 2019**	**After 2019**	**t**	**p**
Bacteria	53.8366533	53.0213562	0.12420784	0.90714232	57.9362804	46.0385277	5.93079572	**0.00405102**
Archaea	6.25499551	6.35264845	-0.0139933	0.98950543	1.37511175	11.0812122	-3.2564603	**0.03118483**
Eukaryota	0.70134512	0.67931355	0.55627367	0.60767315	0.7135128	0.63172842	2.56154585	0.06253564
Viruses	0.33876791	0.32461698	0.12484218	0.90667115	0.25909306	0.58221069	-4.6263903	**0.00983381**
Unclassified	38.8682382	39.6220648	-0.5104975	0.63658503	39.716002	41.666321	-0.8292081	0.45360208

En promedio, la comunidad microbiana está dominada por bacterias del filo *Proteobacteria* (44.32%), seguidas de *Bacteroidetes* (10.66%), *Cianobacterias* (10.49%) y *Chloroflexi* (9.14%) (Figura 3). Las cianobacterias alcanzaron su mayor abundancia en las muestras poco profundas, llegando a ser dominantes en M3 (45.08%), pero sufriendo un gran descenso en M5 y M6 (una media del 1.96%) y en las muestras profundas (una media del 0.18%). *Chloroflexi* mantuvo una abundancia inferior al 10% en la mayoría de las muestras, pero pareció volverse más abundante con la profundidad, alcanzando una abundancia media del 34.26% en D30 y D50. También es importante señalar que *Euryarchaeota* tuvo una abundancia inferior al 0.1% en las muestras superficiales de 2016 a 2018, pero alcanzó una abundancia media del 14.26% en M5 y M6, tras lo cual su abundancia nunca volvió a ser tan baja. Por su parte, la fracción viral fue dominada por lecturas asignadas al orden *Caudovirales*, concretamente a las familias *Siphoviridae* (53.35%), *Myoviridae* (31.48%) y *Podoviridae* (11.92%) (Anexo 1: Supplementary figure 2A).

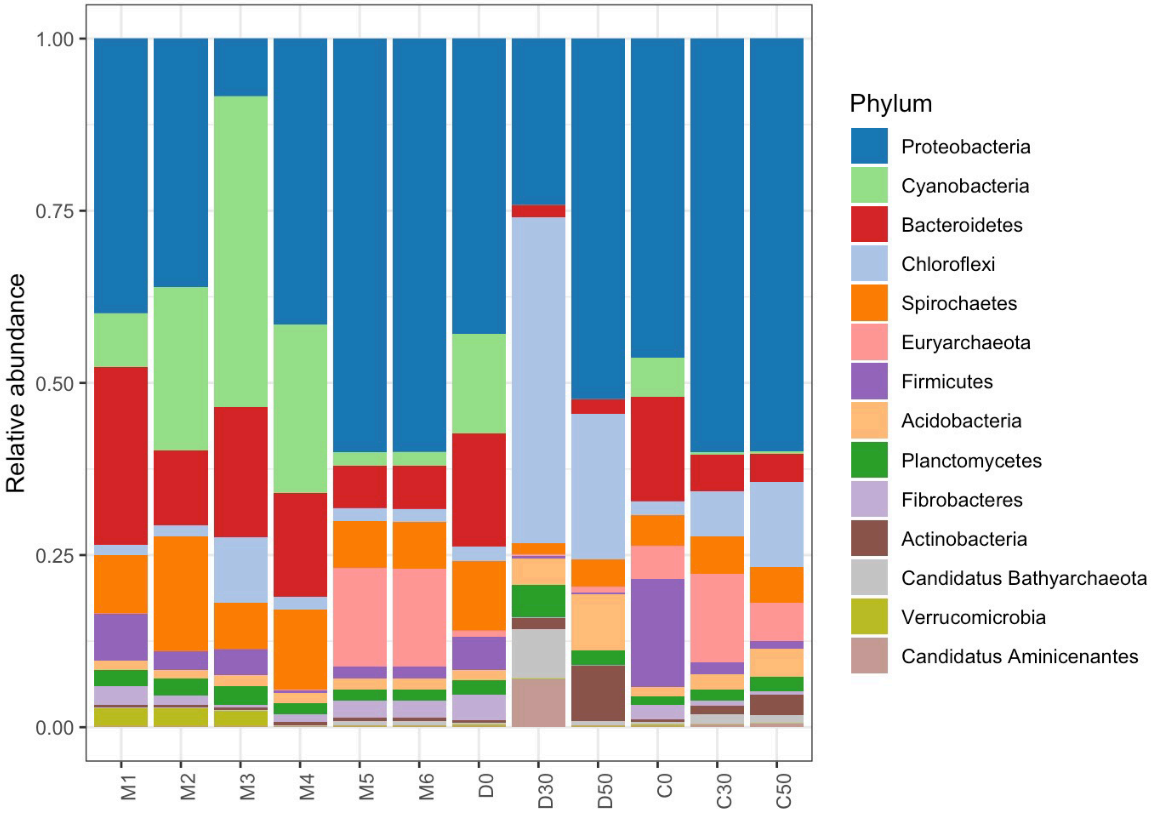


Figura 3. Abundancia relativa de lecturas asignadas a divisiones de bacterias y arqueas en los 12 metagenomas de los Domos del Arqueano.

A nivel de especies virales se pudo observar que a partir de las muestras tomadas en 2019 (M5 y M6) hubo un aumento de las lecturas asignadas a haloarqueavirus que comúnmente se encuentran en otros sitios hipersalinos (Figura 4). En las mismas muestras se produjo una disminución considerable de las lecturas asignadas a *Microviridae* sp., *Circoviridae* sp., *Microvirus* sp., Prokaryotic dsDNA virus sp, uncultured marine phage y *Synechococcus* phage S-SCSM1. De manera similar, en las muestras de 30 y 50 cm de profundidad, se observó un aumento considerable de las lecturas asignadas a virus que infectan Archaeas halófilas y algunos halófagos ambientales sin hospedero conocido.

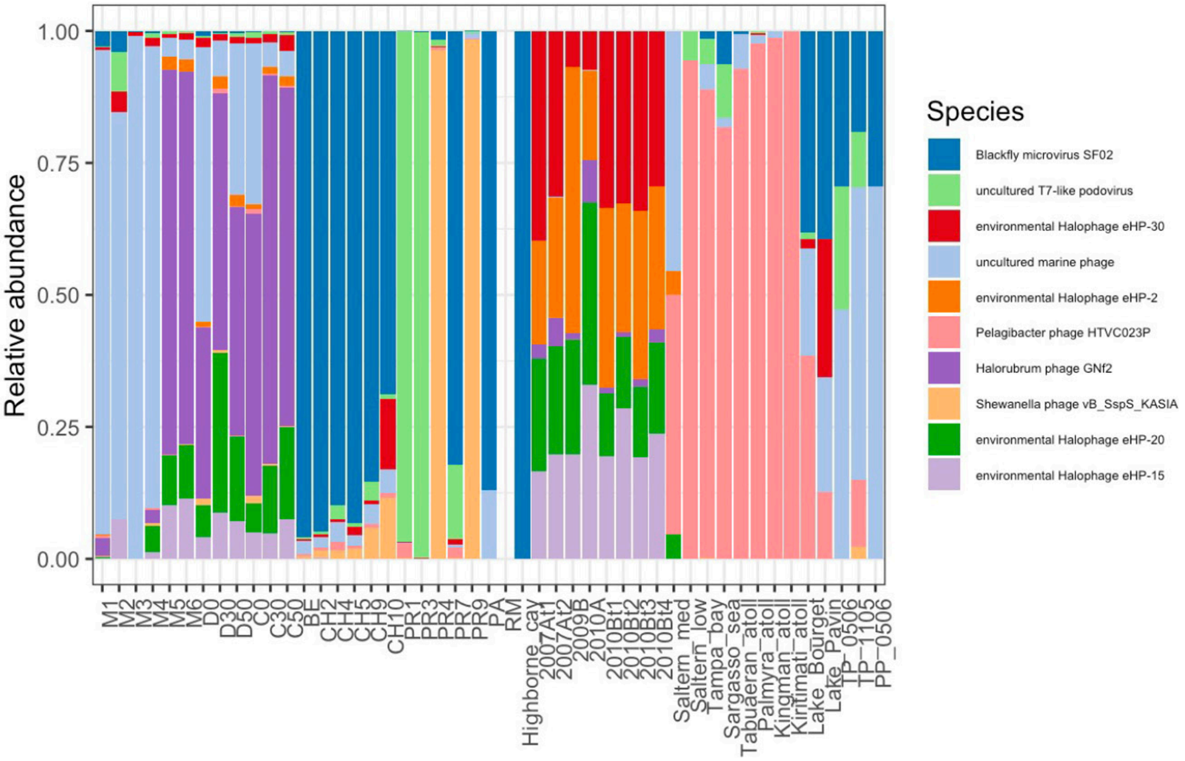


Figura 4. Abundancia relativa de lecturas asignadas a especies virales. Se muestran algunas de las especies más abundantes en los 47 metagenomas. A partir de M5 se observa una gran abundancia de lecturas asignadas a virus de arqueas halófilas, los cuales también se encuentran en otros sitios hipersalinos (2007At1-2010Bt4).

Por último, se observó que las muestras superficiales M5 y M6 tienen abundancias relativas similares a las de muestras profundas (C30, C50, D30 y D50). Este conjunto de muestras se distinguió del resto de las muestras superficiales (M1, M2, M3, M4, C0 y D0) por una mayor abundancia de virus que infectan arqueas halófilas y una menor abundancia de otros virus como los cianófagos.

Comparación de la diversidad viral de AD con la de otros sitios alrededor del mundo

Dentro de AD, el índice de diversidad de Shannon fue mayor para los viromas de 30 y 50 cm de profundidad en comparación con los viromas superficiales de 2016 a 2018 (t-Student, t = 4.44, p = 0.0044) y los superficiales de 2019 y 2020 (t-Student, t = 3.04, p = 0.0228). El índice de Chao1 fue significativamente mayor (t-Student, t = 2.68, p = 0.0368) para los viromas superficiales de 2019 y 2020 en comparación con los viromas superficiales de 2016-2018, mientras que el índice de Simpson fue mayor (t-Student, t = 9.9, p = 6.12e-05) en los viromas de 30 y 50 cm de profundidad en comparación con los viromas superficiales de 2016-2018.

Dado que las muestras superficiales de 2019 (M5 y M6) mostraron un perfil taxonómico más similar al de las muestras de 30 y 50 cm de profundidad (Figura 5), agrupamos las muestras superficiales de 2019 con las muestras de 30 y 50 cm de profundidad (muestras profundas, *sensu lato*) y las comparamos con el resto de muestras superficiales. Tanto el índice de Simpson como el de Shannon presentaron valores más altos para los viromas profundos *sensu lato* en comparación con los superficiales (p = 2.12e-07 y p = 0.0014, respectivamente). No se observaron diferencias entre las temporadas seca y húmeda.

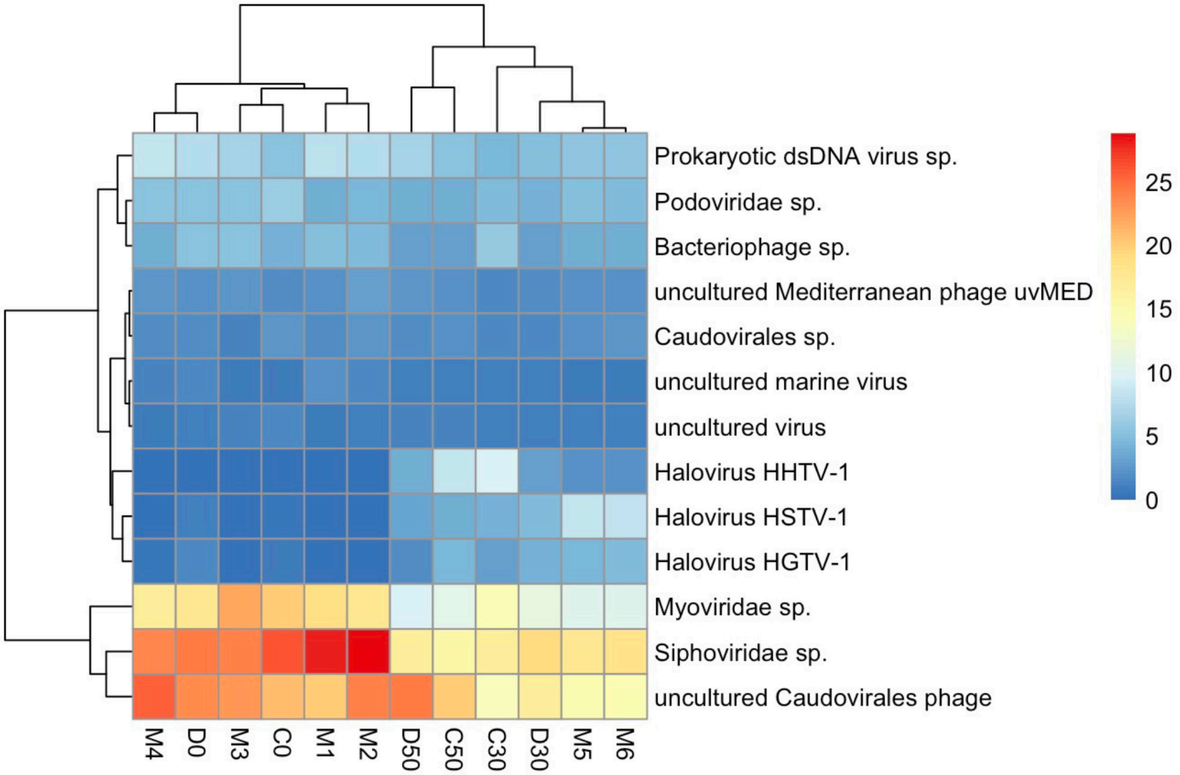


Figura 5. Agrupamiento jerárquico basado en la abundancia relativa de lecturas asignadas a especies virales con al menos 1% de abundancia promedio en los 12 metagenomas de Domos del Arqueano. La mayoría de las muestras superficiales quedan agrupadas, excepto por M5 y M6 que agrupan con las muestras de 30 y 50 cm de profundidad.

Las estimaciones de diversidad alfa mostraron que AD alberga la comunidad viral más diversa entre todos los demás viromas disponibles, incluidas las muestras de CCB, hipersalinas, marinas y de agua dulce (Figura 6). Más concretamente, AD muestra valores del índice de Shannon significativamente más altos que los de los viromas de Churince (Taboada et al., 2018) (t-Student, t = 9.12, p = 3.92e-05), Pozas Rojas (Taboada et al., 2018) (t-Student, t = 9.33, p = 3.38e-05), altamente hipersalinos (Emerson et al., 2012) (t-Student, t = 11.25, p = 2.67e-09) y oceánicos (Angly et al., 2006; Dinsdale et al., 2008b; McDaniel et al., 2008) (t-Student, t = 7.75, p = 8.79e-06).

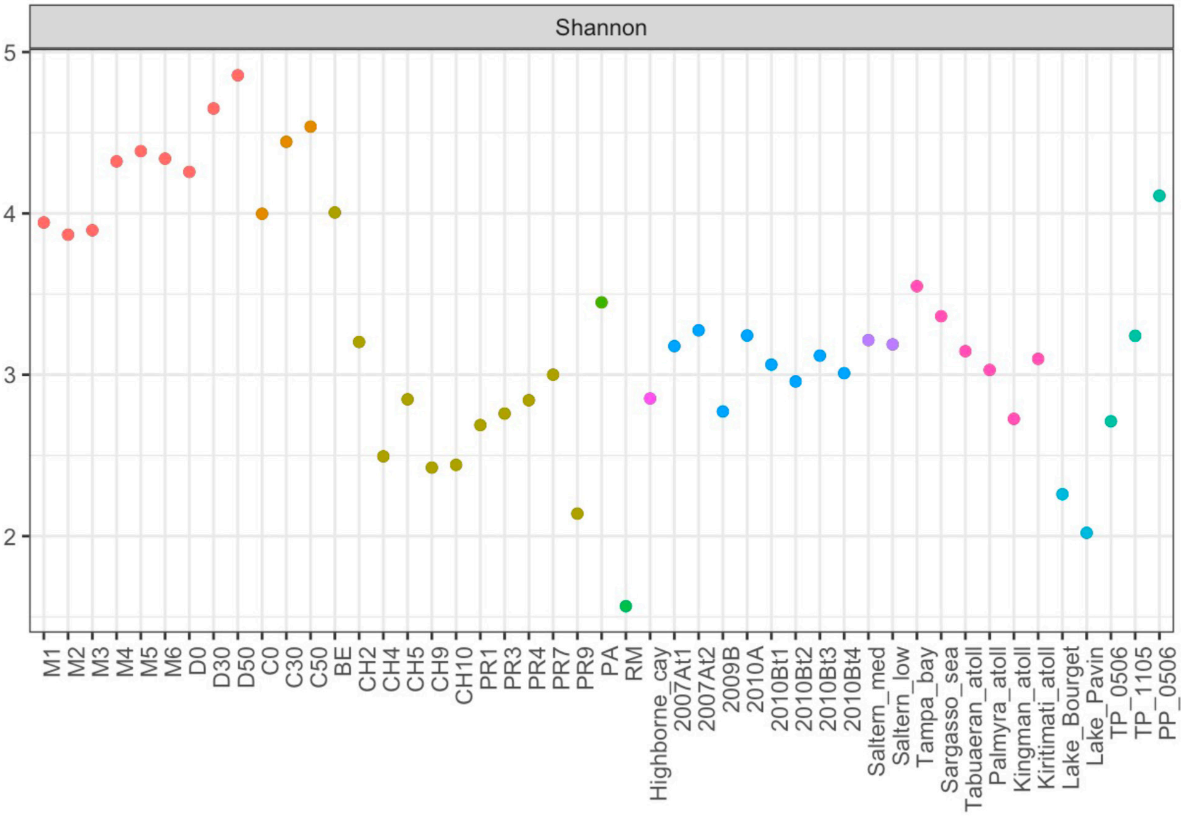


Figura 6. Índice de diversidad de Shannon para los 47 metagenomas comparados. Los metagenoma de Domos del Arqueano muestran la mayor diversidad. Además, se observa una mayor diversidad a mayores profundidades.

La agrupación de Bray-Curtis mostró que los viromas de AD eran más similares a otros viromas de CCB que a los viromas hipersalinos, y que AD formaba una agrupación propia dentro de la cual las muestras se agrupaban por profundidad y no por estaciones. De manera consistente con observaciones anteriores, M5 y M6 se agruparon con los viromas de

30 y 50 cm de profundidad, mientras que los viromas superficiales de 2020 (D0 y C0) se agruparon con los viromas superficiales de 2016 a 2018 (Figura 7).

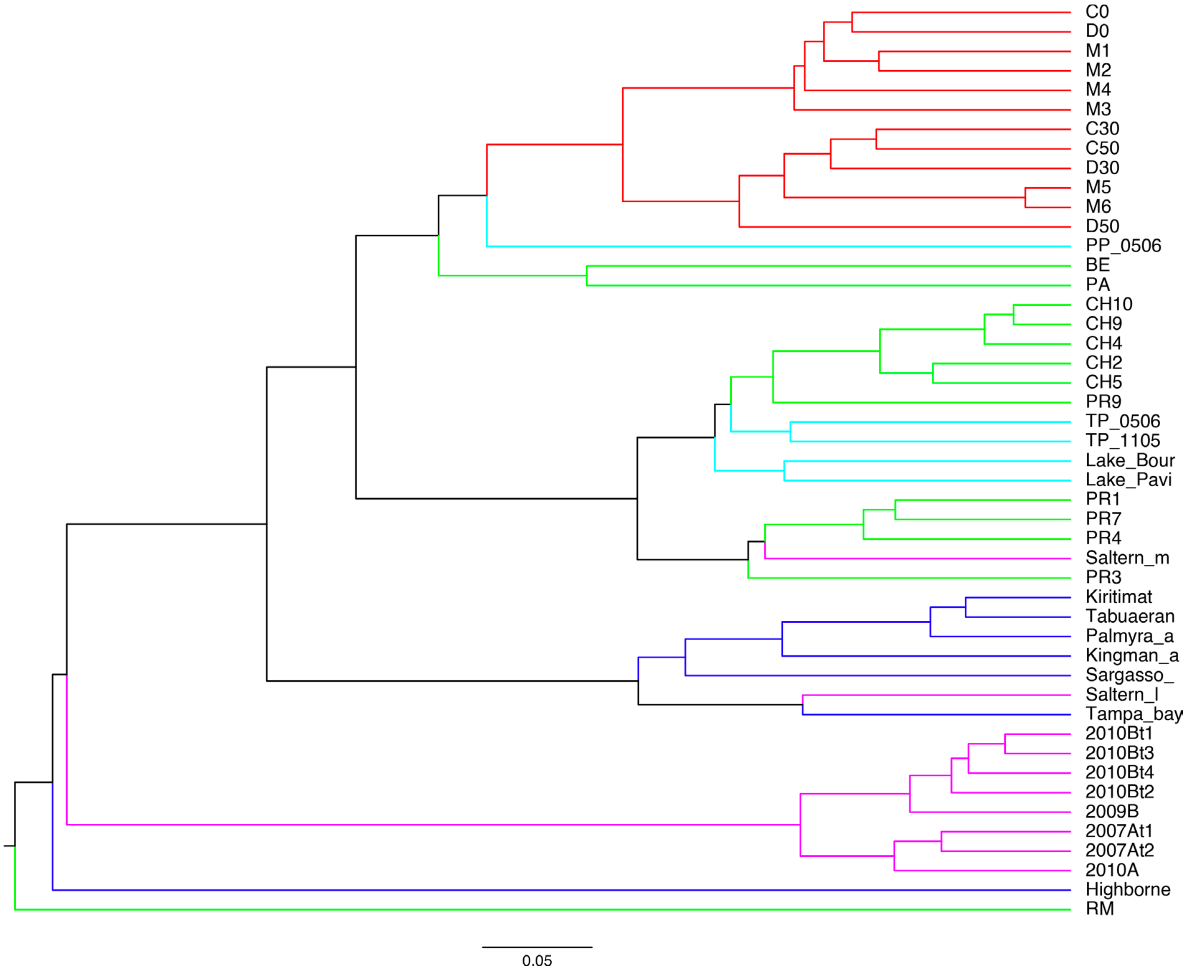


Figura 7. Árbol UPGMA construido a partir de una matriz de disimilitud de Bray-Curtis. SE observa que Domos del Arqueano alberga una comunidad viral única, más similar a la de otros sitios de Cuatro Ciénegas que a la de otros sitios hipersalinos.

Un análisis de redes de similitud de Bray-Curtis reveló que los diferentes viromas se pueden separar en cuatro grupos. El primero conteniendo a todos los viromas de AD, el segundo con todos los viromas derivados de sitios altamente hipersalinos, otro que incluye a todos los viromas marinos y de un sitio ligeramente hipersalino y por último un grupo con viromas de Cuatro Ciénegas (Pozas Rojas y Churince) y de agua dulce. En el mismo análisis se pudo observar que el grupo de viromas de AD se conecta al grupo con otros viromas de Cuatro Ciénegas a través de viromas superficiales (M1, M2, M3, M4, D0 y C0), mientras que la conexión con viromas altamente hipersalinos se da a través de los viromas profundos (C30, C50, D30 y D50) así como M5 y M6 (Figura 8).

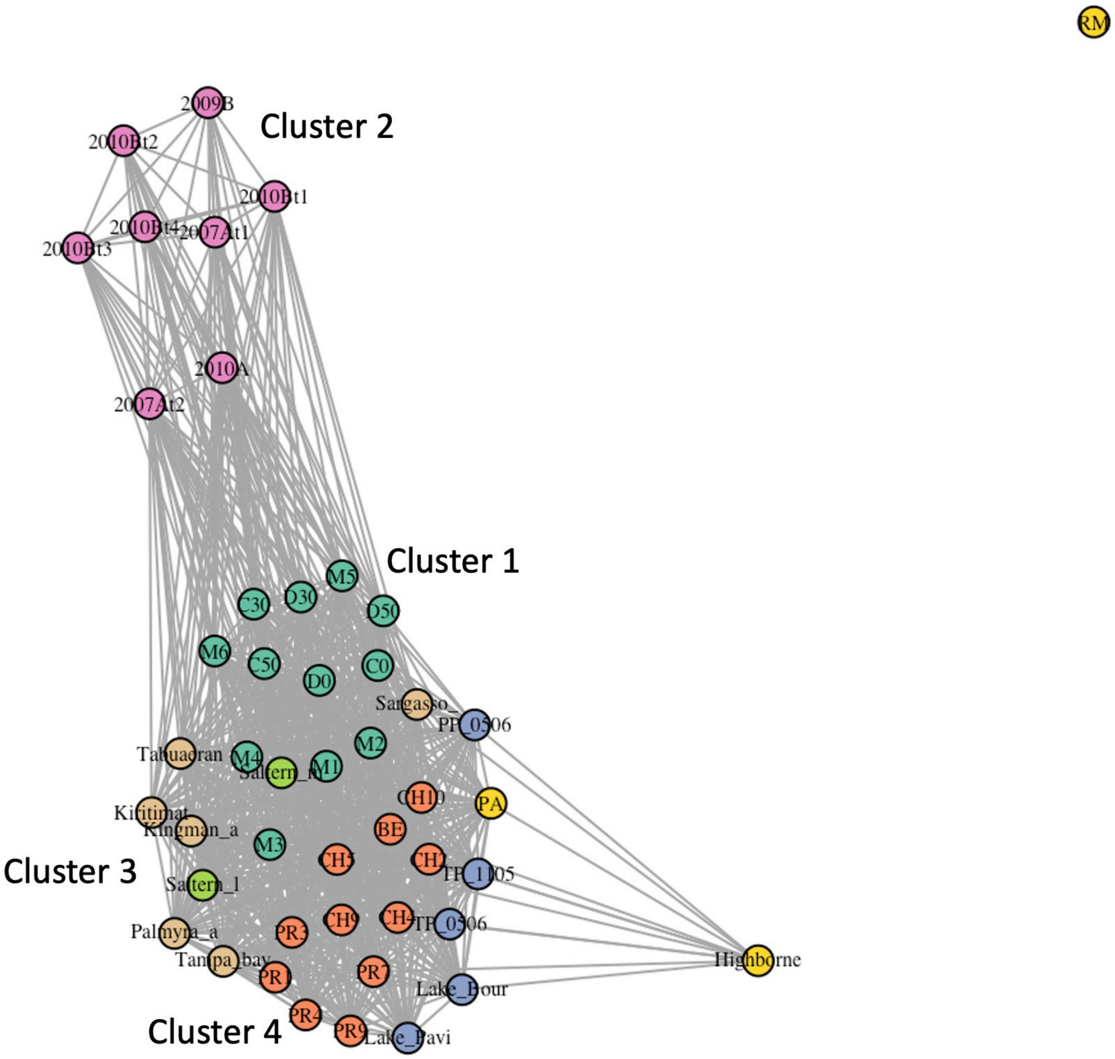


Figura 8. Red de similitud de Bray-Curtis mostrando el 75% de las similitudes más fuertes. Los metagenoams de Domos del Arqueano (verde oscuro) son las que presentan la mayor similitud con los de otros sitios hipersalinos (rosa). La conexión es mediada principalmente por muestras con alta abundancia de virus que infectan arqueas halófilas.

Aunque la comunidad viral de AD no está estrechamente relacionada con las de otros sitios altamente hipersalinos, sí es la que muestra una menor disimilitud contra los mismos. Esto se debe principalmente al incremento en abundancia de virus que infectan arqueas halófilas en M5 y M6 así como en muestras profundas.

Discusión

¿De dónde proviene la diversidad viral de AD?

El concepto de nicho grinnelliano afirma que la estructura de la comunidad está determinada por variables ambientales (Yachi & Loreau, 1999; Soberón, 2007). Se ha demostrado que las comunidades virales hipersalinas siguen patrones globales, de modo que su estructura y diversidad dependen de los cambios en los niveles de salinidad (Roux et al., 2016). Por lo tanto, dada la alta salinidad y las condiciones fluctuantes de las muestras analizadas de los Domos del Arqueano en la Cuenca de Cuatro Ciénegas, esperábamos encontrar una comunidad viral cuya estructura y diversidad, así como su respuesta a los cambios en la salinidad, se asemejaran a las de otros sitios hipersalinos. Sin embargo, encontramos una comunidad dominada por virus pertenecientes al orden *Caudovirales*, como se observa en otros ambientes hipersalinos (Roux et al., 2016) aunque también en otros ambientes templados o extremos (Dávila-Ramos et al., 2019), y una considerable abundancia de lecturas asignadas a haloarqueavirus. Sin embargo, no creemos que ni la estructura de la comunidad ni la diversidad alfa en la comunidad viral de AD estén impulsadas simplemente por las fluctuaciones ambientales. Por ejemplo, ni las muestras de la estación húmeda se agrupan con viromas hipersalinos de salinidad baja o intermedia, ni las muestras de la estación seca se agrupan con otros viromas altamente hipersalinos. En su lugar, los viromas de AD forman un grupo propio, dentro del cual los subgrupos se ordenan por profundidad y no por estación. Además, en contradicción con lo que se ha observado en las comunidades virales de otros lugares hipersalinos (en relación con la disminución de la diversidad cuando la abundancia de haloarqueavirus es elevada) (Roux et al., 2016), en las muestras de AD -en las que hay una mayor abundancia de haloarqueavirus- también se observó una mayor diversidad viral. Por otra parte, en los lagos de sosa etíopes se ha observado una tendencia similar, con un aumento de la diversidad microbiana a mayor salinidad, pH y profundidad (Lanzen et al., 2013).

Si los cambios de temporada no son los principales impulsores de la comunidad, es posible que la dinámica de la comunidad se ajuste a un concepto de nicho Eltoniano, que afirma que la estructura de la comunidad está impulsada por las interacciones (Soberón, 2007; García-Ulloa et al., 2022). Por ejemplo, teniendo en cuenta que los haloarqueavirus desnudos sólo se ven afectados indirectamente por cambios en la salinidad (Luk et al., 2014), que la comunidad viral de la AD está dominada por virus pertenecientes al orden *Caudovirales*, que son virus desnudos, y que recientemente se ha descrito en AD una comunidad microbiana altamente diversa y estacionalmente estable (Madrigal-Trejo et al., 2023), se podría argumentar que la comunidad viral de AD tenderá a permanecer estable mientras la comunidad de hospederos se mantenga igual. Esto puede estar relacionado con la llamada "hipótesis del seguro", que predice que los ecosistemas altamente diversos permanecen funcionalmente estables en entornos cambiantes (Yachi & Loreau, 1999). Además, los organismos que habitan en AD probablemente sean poiquilotróficos, es decir, poliextremófilos adaptados a un entorno sujeto a cambios fisicoquímicos extremos y esporádicos (Gorbushina & Krumbein, 1999). Tal puede ser el caso de los lagos de sosa habitados por microorganismos adaptados tanto a un pH elevado como a la salinidad. Por ejemplo, las comunidades microbianas altamente diversas de los lagos de sosa de la estepa de Kulunda, donde se ha argumentado que las fluctuaciones ambientales (salinidad) promueven el mantenimiento de una alta diversidad (Vavourakis et al., 2016).

Una posible evidencia de interacciones virus-hospedero en AD es la alta abundancia relativa de lecturas asignadas a virus que infectan Archaea, lo cual es consistente con la alta abundancia y diversidad de Archaea reportada en estudios previos (Medina-Chávez et al., 2023; Espinosa-Asuar et al., 2022). Además, el aumento en la abundancia de Archaea al 16% desde 2019 puede estar asociado con un aumento en la abundancia de virus, que a su vez puede contribuir a la alta diversidad y estabilidad de la comunidad microbiana a través de interacciones de "kill the winner" (Wommack & Colwell, 2000; Winter et al., 2010). Sin embargo, se necesitan más análisis para comprobar el alcance y la relevancia de las interacciones virus-hospedero en este sitio.

La mayoría de los ambientes de CCB tienen una baja alcalinidad de carbonatos (Johannesson et al., 2004) y altas concentraciones de Mg^{2+} y Ca^{2+} (Johannesson et al., 2004; Rebollar et al., 2012; Delgado-García et al., 2018) lo que, a pesar de ser atalásico, es muy similar a la composición iónica del agua de mar (Rebollar et al., 2012) donde las concentraciones de Mg^{2+} y Ca^{2+} son mucho mayores que las de carbonatos (Kempe & Kazmierczak, 2011). Dado que la comunidad viral de AD está estrechamente relacionada con la de otros ambientes de CCB podríamos esperar que la composición iónica fuera similar a la del agua de mar. Sin embargo, las mediciones de carbonatos y bicarbonatos en las muestras de AD de 2019 dieron como resultado una mayor alcalinidad total (AT = 2[CO_3^{2-}] + [HCO_3^-]) (de 11,08 mmol/L durante la estación húmeda a 32. 75 mmol/L durante la estación seca) en comparación con el agua de mar (2.33 mmol/L) (Kempe & Kazmierczak, 2011) y otros sitios de CCB (de ~0.5 mmol/L a ~6 mmol/L) (Rebollar et al., 2012), pero no tan alta como en el Lago Van (~150 mmol/L), que es el lago de sosa más grande del mundo (Kempe & Kazmierczak, 2011). Tal alcalinidad puede ser suficiente para especular que el AD es un lago de sosa, sin embargo, esta posibilidad no puede ser confirmada hasta que se haga un análisis aniónico/catiónico completo incluyendo las concentraciones de Mg^{2+} y Ca^{2+}, para probar el criterio de lago de sosa (TA > 2[Mg^{2+}] + 2[Ca^{2+}]) (Kempe & Kazmierczak, 2011).

La elevada salinidad y pH (hasta 9.5 durante la estación húmeda), así como una diversidad y composición de la comunidad similares a las de una muestra de sedimento hipersalino del lago de sosa Hutong Qagan (Mongolia Interior), aumentan la posibilidad de que AD sea un lago de sosa. Sin embargo, los lagos de sosa se consideran los ambientes de pH alto más estables de la Tierra (Jones et al., 1998; Boros & Kolpakova, 2018), debido al efecto amortiguador contra las fuertes variaciones de pH que confiere la alta alcalinidad (Boyd, 2015), lo que contrasta con la caída del pH a 5.5 registrada en AD durante la estación seca (Medina-Chávez et al., 2023), cuando se esperaría que el pH aumentara si se tratara de un verdadero lago de sosa (Kempe & Kazmierczak, 2011). En estos ambientes de alcalinidad limitada, el pH elevado puede ser el resultado de la eliminación neta de CO_2 por la fotosíntesis durante el día. Durante la noche, cuando no hay fotosíntesis, el CO_2 es devuelto al agua a través de la respiración y el pH disminuye de nuevo (Dillon, 2011; Boyd, 2015). Dado que todas las muestras se tomaron durante el día, no podemos saber si este es el caso de AD, sin embargo, podría ser una fuerte posibilidad ya que el pH más alto se observa durante la estación húmeda cuando proliferan las cianobacterias fotosintéticas. Otro proceso que podría explicar la alcalinidad y el pH relativamente más altos en AD en comparación con otros sitios de CCB es la reducción de sulfato, que es un proceso que consume protones llevado a cabo por bacterias reductoras de sulfato. Brevemente, a medida que ocurre la reducción de sulfato, los tapetes de cianobacterias se degradan y la materia orgánica se oxida, lo que resulta en la solubilización de Mg^{2+} de la clorofila y la producción de bicarbonato, respectivamente (Berner et al., 1970; Lyons et al., 1994; Dillon, 2011). Ambos, metabolismos fotosintéticos y sulfato-reductores han sido detectados en AD (Madrigal-Trejo et al., 2023).

Los resultados presentados aquí concuerdan mejor con la hipótesis alternativa de que la comunidad viral de AD será más similar a la de otros sitios CCB debido a su historia geológica compartida y al acuífero profundo. Brevemente, después de la ruptura de Pangea, todo el norte de México estuvo cubierto por un mar somero que comenzó a retroceder a finales del Cretácico debido a la Orogenia Larámide, completando su regresión y el aislamiento de la CCB del Golfo de México con el levantamiento de la Sierra Madre Oriental a principios del Eoceno (Souza et al., 2006; Moreno-Letelier et al., 2012; Souza et al., 2012). Además, estudios isotópicos han demostrado que el agua subterránea del acuífero profundo es una fuente importante para los sistemas acuáticos de CCB (Wolaver et al., 2013), lo que sugiere que el acuífero profundo ha preservado las condiciones de un océano antiguo y ha mantenido linajes microbianos antiguos aislados de sus parientes marinos durante millones de años, tiempo suficiente para permitir el surgimiento de una diversidad microbiana tan grande (Wolaver et al., 2013; Souza et al., 2018).

Aunque se ha demostrado que las comunidades microbianas y virales de CCB tienen altos índices de diversidad α y β (Escalante et al, 2008; Taboada et al., 2018), los análisis de agrupación de las comunidades virales presentados aquí sugieren que las comunidades microbianas de CCB representan un conjunto de comunidades relacionadas. Además, el hecho de que la comunidad vírica de AD forme un conglomerado propio, dentro del cual los subgrupos se ordenan por profundidad y no por estación, que la diversidad aumente a mayor profundidad, y que las muestras superficiales de 2019-2020 presenten rasgos afines a las muestras profundas, sugieren que el acuífero profundo bajo AD alberga una comunidad microbiana altamente diversa que es transportada esporádicamente a la superficie durante eventos de surgencia de agua (tal vez movida por la bolsa magmática en las profundidades de la Sierra San Marcos y Pinos (Wolaver et al., 2013)). Por último, dado que los viromas profundos muestran el mayor porcentaje de lecturas sin clasificar, es probable que la comunidad microbiana del acuífero profundo esté constituida en gran parte por microorganismos aún desconocidos.

Conclusiones

En conjunto, estos resultados muestran que AD alberga una comunidad viral única y muy diversa, rica en haloarqueavirus. Aunque la presencia de haloarqueavirus es única para los viromas de CCB conocidos, la comunidad sigue siendo más similar a las comunidades virales de otros sitios dentro de CCB que a las de otros sitios hipersalinos, excepto cuando se incluyen otros tapetes microbianos hipersalinos.

AD también se distingue de otros sitios hipersalinos por el mantenimiento de una alta diversidad a pesar de los aumentos de salinidad y abundancia de haloarqueavirus. De hecho, la diversidad de AD parece ser mayor que en otros entornos del mundo, a excepción de otros tapetes microbianos hipersalinos y algunos lagos de sosa, que alcanzan niveles de diversidad similares, independientemente de la estación del año.

La singularidad de esta comunidad viral está probablemente relacionada con la gran diversidad de Archaea y las interacciones virus-hospedero que necesitan una mayor exploración para caracterizar completamente la dinámica de la comunidad de este sitio excepcional.

Las similitudes entre los viromas de superficie de 2019-2020 con los viromas de profundidad, que son muy diversos y ricos en haloarqueavirus, apoyan una hipótesis en la que procesos hidrológicos como el afloramiento del acuífero profundo pueden funcionar como un "banco de semillas" con una gran diversidad microbiana.

Capítulo 2. Predicción de hospederos de virus de los Domos del Arqueano

Introducción

Herramientas bioinformáticas para la predicción de virus-hospedero

Las diferentes herramientas para la predicción de pares virus-hospedero pueden agruparse en cinco categorías (Roux et al., 2023): i) métodos dependientes del hospedero, basados en alineamientos; ii) métodos dependientes del hospedero, libres de alineamientos; iii) métodos dependientes de virus, basados en alineamientos; iv) métodos dependientes de virus, libres de alineamientos; y v) métodos integradores (Figura 9).

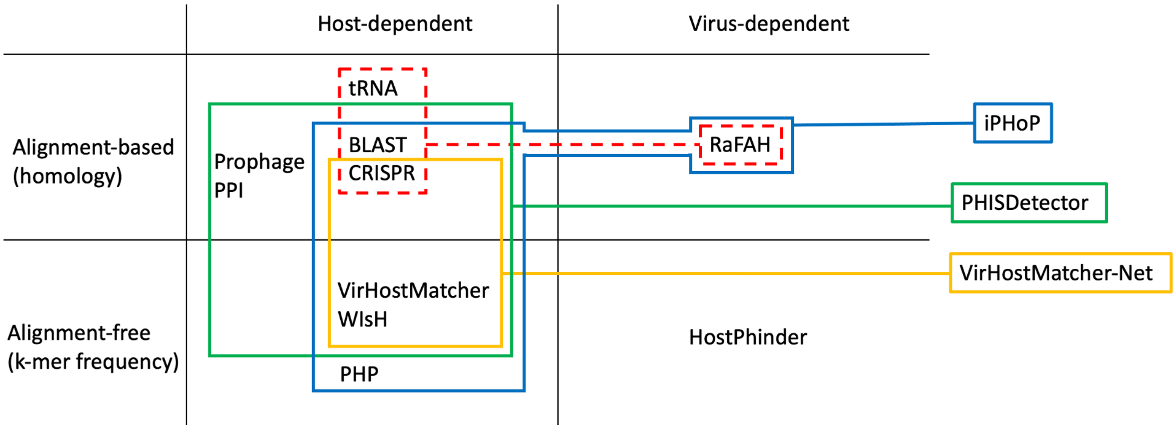


Figura 9. Clasificación de métodos de predicción de virus-hospedero. RaFAH usa métodos dependientes de hospedero, basados en alineamientos, para construir parte de su base de datos de entrenamiento (líneas rojas discontinuas). Los métodos integradores (iPHoP – líneas azules; PHISDetector – líneas verdes; VirHostMatcher-Net -líneas amarillas) tratan de explotar las virtudes de un número diferente de métodos.

Los métodos dependientes del hospedero, basados en alineamientos, incluyen métodos basados en señales de homología, como la búsqueda de homología entre proteínas víricas y bacterianas, tRNAs compartidos, homología entre genomas virales y espaciadores CRISPR, profagos integrados e interacciones proteína-proteína (PPI). Estos métodos son útiles para detectar infecciones recientes, pero tienen el inconveniente de que no todos los virus comparten genes con sus hospederos, lo que tiende a hacerlos precisos, pero con una baja tasa de detección (Edwards et al., 2016). CrisprOpenDB es una herramienta lanzada recientemente que utiliza criterios biológicos para estandarizar las predicciones de hospederos basadas en espaciadores CRISPR con mayor sensibilidad y precisión gracias a su base de datos de >11 millones de espaciadores derivados de >300,000 posibles hospederos (Dion et al., 2021).

Entre los métodos dependientes del hospedero libres de alineamientos se encuentran los basados en la composición de secuencias (por ejemplo, similitud en el uso de codones, similitud en la composición de oligonucleótidos y contenido de GC), que se basan en la noción de que los virus, al ser parásitos genéticos, aproximan su composición de nucleótidos a la del hospedero a lo largo del tiempo. Este mimetismo genómico puede permitir a los virus utilizar los mismos tRNAs para la síntesis de proteínas o eludir los mecanismos de detección y degradación de los ácidos nucleicos extraños. Sin embargo, los virus pueden tener perfiles de secuencia similares de forma independiente, lo que puede dar lugar a una elevada tasa de falsos positivos (Edwards et al., 2016). VirHostMatcher, que evalúa la similitud en la composición de genomas virus-hospedero a través de la distancia d*2 con perfiles de 6-meros (Ahlgren et al., 2017), WIsH, que utiliza perfiles de 8-meros y modelos ocultos de Markov (HMM) (Galiez et al., 2017) y PHP, que utiliza perfiles de 4-meros y un modelo Gaussiano (Lu et al., 2021), son algunos de los métodos libres de alineamiento dependientes del hospedero más conocidos.

En lugar de usar bases de datos de hospederos, los métodos dependientes de virus descansan sobre bases de datos que almacenan virus con hospederos conocidos, de manera tal que los virus de consulta se relacionan con los de las bases de datos bien a través de señales de homología (basados en alineamientos) o bien a través de su similitud en la composición de oligonucleótidos (libres de alineamientos). Por un lado, un programa llamado Random Forest Assignment of Hosts (RaFAH) (Coutinho et al., 2021), utiliza un método dependiente de virus, basado en alineamientos, que construye una parte de su base de datos de entrenamiento a partir de espaciadores CRISPR, la presencia de genes transferidos horizontalmente y tRNAs comunes para, en última instancia, usar aprendizaje de máquina para asociar el virus de consulta a un virus con un hospedero conocido a través de la similitud en el contenido de proteínas. Por otro lado, HostPhinder (Villarroel et al., 2016) es un método sin alineación dependiente de virus que compara perfiles de 16-meros entre virus de consulta y una base de datos de 2,196 fagos con hospederos conocidos.

Por último, los métodos integradores intentan explotar las virtudes de diferentes métodos. Por ejemplo, VirHostMatcher-Net (Wang et al., 2020), que integra métodos dependientes del hospedero basados en alineamientos (espaciadores CRISPR) y métodos dependientes del hospedero libres de alineamientos (VirHostMatcher o WIsH) en un marco de análisis de redes; PHISDetector (Zhou et al., 2022), que integra BLAST (Altschul et al., 1990), espaciadores CRISPR, profagos y análisis PPI a través de un conjunto de enfoques de aprendizaje de máquina; o iPHoP, que utiliza algoritmos de aprendizaje de máquina para calcular puntuaciones conscientes de la taxonomía para BLAST, CRISPR, VirHostMatcher, WIsH y PHP, y las integra con los resultados de RaFAH para obtener una puntuación compuesta final (Roux et al, 2023).

Antecedentes

Virus en los Domos del Arqueano, Cuatro Ciénegas, México

Recientemente se describió la estructura y diversidad de la fracción viral dentro de la comunidad microbiana de la poza hipersalina Domos del Arqueano (AD) en la cuenca de Cuatro Ciénegas (CCB), México (Cisneros-Martínez et al., 2023). En dicho estudio la clasificación taxonómica de las lecturas de secuenciación en escopeta con tecnología Illumina mostró que AD tiene una comunidad viral altamente diversa comparada con las de una variedad de ambientes con diferentes niveles de salinidad distribuidos en varias partes del mundo, incluyendo las de otros sitios de CCB. Tal diversidad incluye una abundancia considerable de virus que infectan arqueas halófilas o haloarqueavirus, la cual incrementa a mayores profundidades. Además de los altos índices de diversidad, AD parece ser un sitio único en el mundo debido a que el incremento en la salinidad y, por lo tanto, en la abundancia de haloarqueavirus, no conduce una disminución en la diversidad como en otros sitios hipersalinos. En cambio, la diversidad en AD parece provenir de eventos de surgencia desde el océano antiguo que yace atrapado en el acuífero profundo de CCB desde hace cientos de millones de años (Cisneros-Martínez et al., 2023).

Justificación

La gran diversidad de virus en AD sugiere que la fracción viral desempeña un papel crucial en el mantenimiento de la diversidad de toda la comunidad microbiana. Sin embargo, para desentrañar la relevancia de los virus en esta comunidad, es importante evaluar la extensión de las interacciones entre los virus y sus hospederos, para lo cual se requiere una mejor caracterización de los mismos que incluya el ensamblaje y anotación de genomas virales a partir de los metagenomas, así como la predicción de los hospederos. La predicción de hospederos es uno de los pasos más importantes en la caracterización de virus ensamblados a partir de metagenomas. Sin embargo, es importante seleccionar las herramientas adecuadas para hacer las predicciones sobre todo si se cuenta con un conjunto de datos altamente diverso que podría incluir múltiples virus nuevos.

Objetivos

General

Caracterizar los virus presentes en AD a partir de la predicción de sus hospederos.

Particulares

1. Ensamblar virus a partir de metagenomas (VEMs) de AD y aplicar las mejores herramientas para realizar la predicción de hospederos.
2. Evaluar la precisión y sensibilidad de una variedad de métodos bioinformáticos para la predicción de virus-hospedero
3. Comparar el desempeño de herramientas que usan grandes bases de datos de referencia con el de herramientas que permiten el uso de bases de datos personalizadas con datos derivados de genomas celulares ensamblados a partir del mismo conjunto de datos metagenómicos (GEMs) que los VEMs.
4. Desarrollar una herramienta de predicción de hospederos basada en espaciadores CRISPR y en los criterios biológicos establecidos por Dion et al. (2021), que permita usar una base de datos personalizada.

Hipótesis

Debido a que AD es un sitio hipersalino en donde se ha descrito una gran diversidad de organismos halófilos y halotolerantes, incluyendo arqueas (Medina-Chávez et al., 2023), así como una gran diversidad de lecturas cortas asignadas a virus de arqueas halófilas (Cisneros-Martínez et al., 2023), se espera que la mayoría de las predicciones de hospederos correspondan a dicho tipo de organismos. Por otro lado, dado que los métodos integradores explotan las virtudes de diferentes herramientas (Wang et al., 2020; Zhou et al., 2022; Roux et al, 2023), podría esperarse que estos mostraran el mejor desempeño. Sin embargo, es posible que la predicción de hospederos de virus altamente diversos (y posiblemente nuevos) como los de AD, sea más precisa cuando se usa una base de datos personalizada (en donde los datos asociados a los posibles hospederos provengan del mismo conjunto de datos metagenómicos que los VEMs) que cuando se usa una base de datos de virus u hospederos conocidos, por extensa que esta sea.

Materiales y métodos

Evaluación comparativa de herramientas bioinformáticas para la predicción virus-hospedero

Los genomas de virus y bacterias se seleccionaron descargando tres listas: i) genomas virales completos del NCBI (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi) filtrados por hospedero "bacteria"; ii) informe tabular de virus-hostDB (https://www.genome.jp/ftp/db/virushostdb/); iii) catálogo del RefSeq release 217 (https://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/), filtrado por molécula genómica completa, no plásmido. Se estableció un vínculo entre las tres tablas, de modo que si el código de acceso del genoma viral tenía una coincidencia en la virus-hostDB, se comprobaba si el taxón del hospedero tenía una coincidencia en el catálogo RefSeq (para un virus dado con hospedero conocido, comprobar si el hospedero tiene un genoma completo). Se descargaron del NCBI un total de 1,029 genomas de fagos y 133 genomas bacterianos, que juntos forman 1,046 pares virus-hospedero con genomas completos.

El rendimiento de las herramientas de predicción virus-hospedero se evaluó a nivel de género. Se elaboró un script personalizado (https://github.com/AleCisMar/CrisprCustomDB/blob/main/benchmarking/compare_real-estimated.pl) para comparar los pares estimados de virus (código de acceso)-hospedero (género) con los pares reales de virus (código de acceso)-hospedero (género) con el fin de obtener los verdaderos positivos (TP) y los falsos positivos (FP) de cada herramienta de predicción. La precisión, la sensibilidad y la puntuación F1 se calcularon del siguiente modo:

Precisión (Positive Predictive Value):

$$PPV = \frac{TP}{TP + FP}$$

Sensibilidad (True Positive Rate):

$$TPR = \frac{TP}{TP + FN}$$

score-F1:

$$F1_{score} = 2\left(\frac{PPV * TPR}{PPV + TPR}\right)$$

Para ejecutar cada uno de los programas, se siguieron las instrucciones proporcionadas por los desarrolladores. Además, se desarrolló un script de perl inspirado en CrisprOpenDB (CrisprCustomDB disponible en https://github.com/AleCisMar/CrisprCustomDB). CrisprCustomDB utiliza los criterios de asignación de hospederos de Dion et al. (2021). Estos son: (i) hospedero si los espaciadores tienen un máximo de 2 *mismatches* con el genoma viral; (ii) si hay múltiples candidatos que cumplen el criterio 1, se selecciona el que tiene más espaciadores alineados con diferentes regiones del genoma viral; (iii) si hay múltiples candidatos que cumplen el criterio 2, se selecciona el que tiene el espaciador más cercano al extremo 5' en el arreglo y; (iv) si hay múltiples candidatos que cumplen el criterio 3 se asigna el rango taxonómico común.

Para probar el rendimiento de CrisprCustomDB en la base de datos personalizada (133 genomas bacterianos completos), los espaciadores CRISPR se obtuvieron con la herramienta CRISPRDetect (Biswas et al., 2016) con los parámetros recomendados por Dion et al. (2021). CRISPRDetect detecta los arreglos CRISPR buscando pares de secuencias repetidas separadas por una secuencia espaciadora, las distingue de repeticiones en tándem, extiende el arreglo, y realiza una predicción de la dirección del arreglo, la cual es necesaria para la implementación del tercer criterio de CrisprCustomDB. Dado que sólo se encontraron 1,349 espaciadores en 40 de los 133 genomas bacterianos (30%) se realizaron dos cálculos de sensibilidad, uno teniendo en cuenta los 1,046 pares virus-hospedero y otro teniendo en cuenta sólo el número máximo de pares virus-hospedero (261) que se pueden obtener dado el número de posibles hospederos con espaciadores (40). CrisprCustomDB, VirHostMatcher, WIsH y PHP se ejecutaron en la base de datos personalizada (1,029 genomas de fagos y 133 genomas bacterianos). Dado que HostPhinder, CrisprOpenDB, VirHostMatcher-Net y RaFAH se basan en grandes bases de datos de referencia, sólo se utilizaron como entrada los 1,029 genomas de fagos. Para PHP se realizó una estimación adicional utilizando la base de datos de referencia con 60,105 hospederos potenciales proporcionada por los autores. Para VirHostMatcher se hicieron dos estimaciones, ambas con puntuación ≤ 0.25. La primera seleccionando el hospedero más frecuente dentro del top 30 con los perfiles más similares y la segunda seleccionando el hospedero más frecuente dentro del top 5 con los perfiles más similares. Para VirHostMatcher-Net también se realizaron dos estimaciones. Una sin restricción de puntuación y la otra limitada a predicciones con puntuación > 0.95.

Procesamiento de lecturas y ensamblaje de VEMs

La calidad de las lecturas se evaluó con FastQC v0.11.9 (Andrews, 2010). La eliminación de adaptadores y el filtrado de la calidad se realizaron con Trimmomatic v0.39 (Bolger et al., 2014) utilizando una ventana deslizante de 4 pares de bases que excluía las lecturas con una calidad media inferior a 30 y con menos de 20 nucleótidos. Las lecturas limpias se ensamblaron con SPAdes 3.15.2 (Antipov et al., 2020) utilizando la opción --metaviral. Se utilizaron los scripts viralVerify y viralComplete (incluidos en el paquete SPAdes) para verificar que los contigs ensamblados correspondían a genomas virales y para evaluar la completitud del genoma, respectivamente. El resultado fueron 87 contigs virales completos. Se comprobó la circularidad de los contigs virales. Cuando fue necesario, se ajustó la posición de las secuencias antes de la predicción y anotación de genes con la ayuda de scripts personalizados (disponibles en https://github.com/AleCisMar/GenomicTools) que hacen uso de BLAST (Altschul et al., 1990), EMBOSS (Rice et al., 2000), Prodigal (Hyatt et al., 2010) y HMMER (Eddy, 2011).

Procesamiento de las lecturas, ensamblaje y asignación taxonómica de los GEMs de bacterias y arqueas

La calidad de las lecturas crudas se evaluó con FastQC (v0.11.8) (Andrews, 2010) y se filtró con Trimmomatic (v0.39) (Bolger et al., 2014). A continuación, las lecturas se ensamblaron con MetaSPAdes (v3.15.3) (Nurk et al., 2017) y los contigs obtenidos en el ensamblaje se utilizaron para realizar binning o clustering de lecturas con MaxBin2 (v2.2.7) (Wu et al., 2015) y MetaBat2 (v2.12.1) (Kang et al., 2019). Se utilizó el software Binning refiner (v1.4.2) (Song & Thomas, 2017) para reducir el porcentaje de contaminación en los bins (Song & Thomas, 2017). La integridad de los GEMs se evaluó utilizando CheckM (v1.1.3) (Parks et al., 2015) con la configuración predeterminada. Los GEMs se filtraron utilizando los siguientes criterios: > 70 % de integridad y < 10 % de contaminación, dando como resultado 940 GEMs. Para la asignación taxonómica y la colocación de los GEMs en el árbol filogenético de la vida, se utilizó el programa GTDB-tk (v1.6.0) (Chaumeil et al., 2019), que identifica 122 y 120 genes marcadores de arqueas y bacterias, respectivamente, utilizando HMMER (Eddy, 2011). En resumen, los genomas se asignan al dominio con la mayor proporción de genes marcadores identificados. Los marcadores específicos de dominio seleccionados se alinean con HMMER, se concatenan en una única alineación de secuencias múltiples y se recortan con la máscara de columna de ~5,000 Bacteria o Archaea utilizada por GTDB (Chaumeil et al., 2019).

Implementación de herramientas de predicción virus-hospedero en datos metagenómicos

Tras la clasificación taxonómica de los GEMs, se realizó la predicción de las secuencias espaciadoras de los arreglos CRISPR encontrados en los GEMs con el programa CRISPRCasTyper (v 1.3.0) (Russel et al. 2020) utilizando los siguientes parámetros: cctyper -t 4 --prodigal single -circular. Se encontró un total de 2,660 espaciadores en el 19.15% de los GEMs. Los 87 VEMs se cotejaron con la base de datos de espaciadores con blastn (Altschul et al., 1990) permitiendo un máximo de 2 *mismatches*. Para implementar CrisprCustomDB se predijeron los espaciadores con la herramienta CRISPRDetect (Biswas et al., 2016) utilizando -array_quality_score_cutoff de 3 como se recomienda para los archivos FASTA. El resultado fueron 1,062 espaciadores predichos en el 11.7% de los GEMs. Los 87 VEMs también se ejecutaron con CrisprOpenDB, que utiliza una base de datos de 11,674,395 espaciadores (Dion et al., 2021). También se realizaron predicciones virus-hospedero con RaFAH (Coutinho et al., 2021) y PHP (Lu et al., 2021). Para PHP se utilizaron los 940 GEMs

Resultados

Desempeño de las herramientas bioinformáticas para la predicción de virus-hospedero

Las tres herramientas con mejor rendimiento (F1_score = media armónica entre la precisión y la sensibilidad) en el conjunto de datos de genomas completos de bacterias y fagos, fueron RaFAH, PHP y VirHostMatcher-Net (Tabla 3). A continuación, se situaron WIsH, VirHostMatcher, CrisprOpenDB, HostPhinder y CrisprCustomDB. CrisprOpenDB realizó 392 predicciones, de las cuales 259 fueron estimaciones correctas. Estos resultados se traducen en una sensibilidad del 24.76%, una precisión del 66.07% y un F1_score del 36.02%. Por otro lado, CrisprCustomDB sólo predijo 28 pares, todos ellos correctos (precisión = 100%). Teniendo en cuenta que realizamos una evaluación comparativa con 1,046 pares virus-hospedero, CrisprCustomDB alcanzó una sensibilidad de sólo el 2.68% y un F1_score del 5.21%. Es importante señalar que sólo se identificaron espaciadores en el 30% de los genomas bacterianos de la base de datos personalizada utilizando CRISPRDetect. En consecuencia, el número máximo de pares predichos fue de 261. Esta consideración aumenta la sensibilidad al 10.73% y la puntuación F1 al 19.38% (Figura 10).

Tabla 3. Precisión, sensibilidad, y F1_score de diferentes herramientas de predicción de virus-hospedero.

Software	Actual virus-host pairs	Predicted pairs	NA	True positive	False positive	False negative	Precision	Sensitivity	F1_score
CrisprCustomDB	1046	28	1018	28	0	1018	1	0.0268	0.0521
CrisprCustomDB*	261	28	233	28	0	233	1	0.1073	0.1938
HostPhinder	1046	1044	2	367	677	679	0.3515	0.3509	0.3512
CrisprOpenDB	1046	392	654	259	133	787	0.6607	0.2476	0.3602
VirHostMatcher[†]	1046	743	303	459	284	587	0.6178	0.4388	0.5131
PHP[§]	1046	1001	45	550	451	496	0.5495	0.5258	0.5374

VirHostMatcher[¶]	1046	638	408	604	34	442	0.9467	0.5774	0.7173
WisH	1046	1046	0	794	252	252	0.7591	0.7591	0.7591
VirHostMatcher-Net**	1046	903	143	829	74	217	0.9181	0.7925	0.8507
VirHostMatcher-Net	1046	1046	0	921	125	125	0.8805	0.8805	0.8805
PHP	1046	1046	0	952	94	94	0.9101	0.9101	0.9101
RaFAH	1046	1046	0	1001	45	45	0.957	0.957	0.957

*Sensibilidad calculada a partir de 261 pares posibles dado el número de hospederos con espaciadores.

[†] Predicción usando score ≤ 0.25 y seleccionando el hospedero más frecuente dentro del top 30.

[§] Usando la base de datos de referencia de PHP con 60,105 genomas de procariontes.

[¶] Predicción usando score ≤ 0.25 y seleccionando el hospedero más frecuente dentro del top 5.

**Predicción usando score > 0.95.

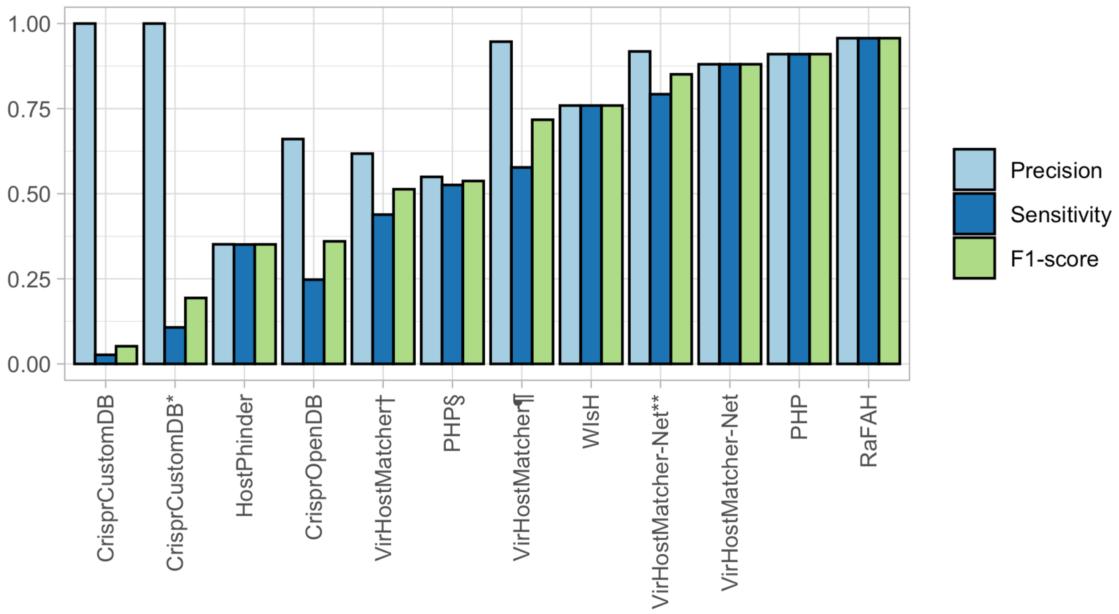


Figura 10. Precisión, sensibilidad, y F1_score de diferentes herramientas para predicción de virus-hospedero. Para CrisprCustomDB*, la sensibilidad se estimó considerando 261 pares posibles. VirHostMatcher† se ejecutó con un score ≤ 0.25 y seleccionando al hospedero más frecuente dentro del top 30. VirHostMatcher¶ se ejecutó con los mismos parámetros pero seleccionando al hospedero más frecuente dentro del top 5. PHP§ se ejecutó contra una base de datos de 60,105 posibles hospederos. Para VirHostMatcher-Net**, solo se usaron predicciones con score > 0.95.

De los métodos libres de alineamientos, HostPhinder obtuvo el peor desempeño, con una sensibilidad del 35.09%, una precisión del 35.15% y una puntuación F1_score del 35.12%. VirHostMatcher se ejecutó con dos criterios diferentes: i) seleccionando el hospedero más frecuente entre los treinta primeros y; ii) seleccionando el hospedero más frecuente entre los cinco primeros. Al utilizar el primer criterio, VirHostMatcher generó más predicciones (743 frente a 638) y produjo más falsos positivos (284 frente a 34). Como resultado, obtuvo menor sensibilidad (43.88% frente a 57.74%), precisión (61.78% frente a 94.67%) y F1_score (51.31% frente a 71.73%).

WIsH y PHP fueron los mejores predictores sin alineamientos dependientes de hospedero, alcanzando el máximo número de pares (1,046). WIsH demostró una sensibilidad, precisión y F1_score del 75.91%, mientras que PHP obtuvo una sensibilidad, precisión y F1_score del 91.01%. PHP también fue probado contra una base de datos de referencia con 60,105 posibles hospederos proporcionada por los desarrolladores. Sin embargo, esta prueba dio como resultado menos predicciones (1,001) y sensibilidad (52.58%), precisión (54.95%) y F1_score (53.74%) más bajos.

VirHostMatcher-Net se ejecutó utilizando dos aproximaciones: en primer lugar, estableciendo una puntuación > 0.95 para predicciones válidas y, en segundo lugar, sin ninguna restricción de puntuación. Restringir la asignación final de hospederos a las predicciones con puntuaciones más altas dio como resultado una mayor precisión (91.81% frente a 88.05%) a costa de una menor sensibilidad (79.25% frente a 88.05%) y, como consecuencia, una puntuación F1 más baja (85.07% frente a 88.05%). Por su parte, RaFAH alcanzó una precisión, sensibilidad y puntuación F1 del 95.70%, lo que lo convierte en la herramienta con mejor desempeño de todas.

Predicción de hospederos de VEMs de Domos del Arqueano

Se siguieron dos aproximaciones para predecir hospederos de VEMs de Domos del Arqueano con base en espaciadores CRISPR derivados de GEMs del mismo conjunto de datos metagenómicos. Por un lado, se realizó una búsqueda con Blastn utilizando 2,660 espaciadores detectados con CRISPRCasTyper, con un máximo de 2 discordancias como único criterio. Por otro lado, se usó CrisprCustomDB, utilizando 1,062 espaciadores detectados con CRISPRDetect, para resolver asignaciones problemáticas de hospederos.

Además, realizamos predicciones utilizando CrisprOpenDB, PHP y RaFAH, ya que estas herramientas demostraron un rendimiento superior en sus respectivas categorías. Dado que HostPhinder mostró un rendimiento inferior al resto de métodos sin alineación, y que PHP y RaFAH superaron a VirHostMatcher-Net, no realizamos predicciones con estas herramientas. Mientras que PHP se ejecutó en los GEMs de Domos del Arqueano, CrisprOpenDB y RaFAH sólo requirieron los VEMs, ya que se basaron en sus bases de datos de referencia.

Usando espaciadores CRISPR con Blastn y sin criterios adicionales, se obtuvieron ocho predicciones. La mitad de los VEMs (C50N1L42, C0N5L506, M1N5L607 y C9N1L394) se asignaron a hospederos del filo *Desulfobacterota*. La otra mitad fueron asociados a bacterias de la clase *Gammaproteobacteria*. Se predijo que dos VEMs (M5N2L438 y M6N1L439) infectan bacterias del género *Halorhodospira*. Un contig, el M4NL642, se asignó al género *Halochromatium*, y el contig C30N1L64, a dos posibles hospederos: *Thiohalorhabdus* o *Thiohalospira* (Tabla 4).

Tabla 4. Predicciones de hospederos designadas como confiables con base en la consistencia entre métodos o entre el ambiente de origen y la biología del hospedero predicho (taxonomía, hábitat, estilo de vida y metabolismo).

Contig	CRISPR	CrisprCustomDB	CrisprOpenDB	PHP	RaFAH	Supporting evidence	Reference
C50N1L42	*Desulfovibrionales;Desulfovermiculus*	NA	NA	*Desulfovibrionales;Desulfohalobiaceae*	*Desulfovibrionales;Desulfovibrio*	Halophilic; sulfate-reducing	[16, 17, 35]
M5N2L438	*Halorhodospira*	*Halorhodospira*	NA	*Halorhodospira*	*Pseudomonas*	Halophilic	[36]
M6N1L439	*Halorhodospira*	*Halorhodospira*	NA	*Halorhodospira*	*Pseudomonas*	Halophilic	[36]
C30N1L64	*Thiohalorhabdus/ Thiohalospira*	*Thiohalorhabdus**	NA	*Thiohalorhabdus*	*Vibrio*	Halophilic; sulfur-oxidizing	[35, 37]
C0N5L506	*Desulfobacterales*	NA	NA	*Desulfobacterales*	*Clostridium*	Sulfate-reducing	[16]
M1N5L607	*Desulfobacterales*	*Desulfobacterales*	NA	NA	*Pseudoalteromonas*	Sulfate-reducing	[16]
C0N1L394	*Desulfohalobiaceae;Desulfovermiculus*	NA	NA	*Desulfohalobiaceae*	*Bacteroides*	Halophilic; sulfate-reducing	[16, 17, 35]
M4N1L642	*Halochromatium*	*Halochromatium*	*Thiobacillus*	NA	*Kingella*	Halophilic	[38]
C0N2L458	NA	NA	*Gammaproteobacteria;Halomonas*	*Gammaproteobacteria;Halochromatium*	*Thauera*	Halophilic	[35, 38]
M5N6L415	NA	NA	NA	*Archaea;Hadarchaeia*	*Archaea;Haloarcula*	Archaea	[16]

M6N2L524	NA	NA	NA	*Halobacteriales; Halorubrum*	*Halobacteriales; Haloarcula*	Halophilic archaea	[16]
M1N1L790	NA	NA	*Halanaerobium*	NA	*Clostridium*	Halophilic; thiosulfate-reducing	[16, 17, 35]
D30N111L	NA	NA	NA	*Archaeoglobaceae*	*Pseudomonas*	Archaea	[16]
D30N2L48	NA	NA	NA	*Bathyarchaeia*	*Veillonella*	Archaea	[16]
D30N115L	NA	NA	NA	*Nanoarchaeia*	*Bacillus*	Archaea	[16]
M4N1L424	NA	NA	NA	*Dichotomicrobium*	*Parabacteroides*	Thermohalophilic	[39]
D30N1L56	NA	NA	NA	*Aminicenantaceae*	*Clostridium*	Deep marine sediments	[40]
M3N8L364	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
M1N5L608	NA	NA	NA	*Anaerolineae*	*Fusobacterium*	Deep marine sediments	[41]
M5N3L645	NA	NA	NA	*Anaerolineae*	*Kingella*	Deep marine sediments	[41]
C50N2L80	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
D50N2L80	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
M5N8L404	NA	NA	NA	*Bipolaricaulia*	*Leptotrichia*	Hypersaline sediments	[42]
M6N4L404	NA	NA	NA	*Bipolaricaulia*	*Leptotrichia*	Hypersaline sediments	[42]
D30N50L3	NA	NA	NA	*Bipolaricaulia*	*Vibrio*	Hypersaline sediments	[42]
C50N1L90	NA	NA	NA	*Chitinivibrionales*	*Prevotella*	Haloalkaliphilic	[43]
M1N25L46	NA	NA	NA	*Chitinivibrionales*	*Chlamydia*	Haloalkaliphilic	[43]
D30N6L39	NA	NA	NA	*Chitinivibrionales*	*Porphyrobacter*	Haloalkaliphilic	[43]
M1N22L26	NA	NA	NA	*Chitinivibrionales*	*Pseudomonas*	Haloalkaliphilic	[43]
M5N4L592	NA	NA	NA	*Halothiobacillaceae*	*Faecalibacterium*	Halotolerant; halophilic	[44]
M6N2L592	NA	NA	NA	*Halothiobacillaceae*	*Faecalibacterium*	Halotolerant; halophilic	[44]
M5N7L416	NA	NA	NA	*Wenzhouxiangella*	*Burkholderia*	Haloalkaliphilic	[45]
M6N3L417	NA	NA	NA	*Wenzhouxiangella*	*Burkholderia*	Haloalkaliphilic	[45]
M1N1L521	NA	NA	NA	*Halofilum*	*Vibrio*	Marine solar saltern	[46]
M5N28L50	NA	NA	NA	*Gemmatimonadetes*	*Haloarcula*	Halophilic archaea	[16]
M6N4L511	NA	NA	NA	*Gemmatimonadetes*	*Haloarcula*	Halophilic archaea	[16]
C0N2L195	NA	NA	NA	*Halanaerobiales*	*Alistipes*	Halophilic; thiosulfate-reducing	[16, 17, 35]
M4N3L527	NA	NA	NA	*Phycisphaerales*	*Pseudomonas*	Marine	[47]
M1N3L461	NA	NA	NA	*Rhodothermales*	*Alistipes*	Thermohalophilic; haloalkaliphilic	[48, 49]

D30N26L5	NA	NA	NA	*Petrotogales*	*Fusobacterium*	Thermophilic	[50]
C0N1L567	NA	NA	NA	*Halanaerobiales*	*Clostridium*	Halophilic	[51]
C30N1L45	NA	NA	NA	NA	*Salinispora*	Marine sediments	[52]
M1N6L535	NA	NA	NA	NA	*Desulfotomaculum*	Thermophilic; sulfate-reducing	[53]
M1N8L483	NA	NA	NA	NA	*Thermus*	Thermophilic	[54]
M5N19L71	NA	NA	NA	NA	*Caulobacter*	Oligotrophic	[55]
M6N6L714	NA	NA	NA	NA	*Caulobacter*	Oligotrophic	[55]

Se muestran predicciones de hospederos para 46 VEMs de Domos del Arqueano. La lista completa (S4_File) así como las referencias se encuentran en el artículo. Las predicciones confiables se muestran subrayadas. En caso de aplicar, el rango taxonómico más bajo en común, y el rango taxonómico más bajo alcanzado por cada herramienta, se muestran separados por ";".

*Asignado usando el criterio 3: Múltiples hospederos cuyos espaciadores alinean con el mismo número de regiones del genoma viral. Se selecciona el hospedero con el espaciador más cercano al extremo 5'.

CrisprCustomDB realizó cinco predicciones, todas ellas coherentes con las obtenidas con la aproximación anterior. Estas incluyeron una de las *Desulfobacterota* y tres *Gammaproteobacteria*. La quinta predicción asignó el contig C30N1L64 a *Thiohalorhabdus* por ser el hospedero con el espaciador más cercano al extremo 5', resolviendo el problema de dos posibles hospederos. Por su parte, CrisprOpenDB realizó cinco predicciones. El único contig para el que los tres métodos basados en espaciadores CRISPR hicieron una predicción es el contig M4N1L642. Sin embargo, método ordinario y CrisprCustomDB predijeron que infecta a la proteobacteria *Halochromatium*, mientras que CrisprOpenDB predijo que el hospedero del fago pertenece al género *Thiobacillus*. Por último, CrisprOpenDB predijo que el contig C0N2L458 infecta a alguna *Gammaproteobacteria* del género *Halomonas*.

PHP hizo 54 predicciones. Tres de los contigs asignados a *Desulfobacterota* por la aproximación ordinaria basada en espaciadores CRISPR (C50N1L42, C0N5L506, y C9N1L394) fueron igualmente asignados por PHP. Además, tres contigs asignados a *Proteobacteria* tanto en el enfoque CRISPR ordinario como en el CrisprCustomDB fueron asignados independientemente por PHP. Esta concordancia incluye los contigs M5N2L438 y M6N1L439, asignados a *Halorhodospira*, y el contig C30N1L64, que también fue asignado a *Thiohalorhabdus*. Además, PHP coincidió con CrisprOpenDB en la asignación de hospedero para el contig C0N2L458 a nivel de clase, pero sugirió que infecta a bacterias del género *Halochromatium* en lugar de *Halomonas*.

RaFAH produjo 87 predicciones, de las cuales sólo tres fueron apoyadas por los otros métodos. Estas incluyeron la asignación de hospedero para el contig C50N1L42 (*Desulfobacterota*), que es consistente con el enfoque CRISPR ordinario y PHP, y la asignación de hospedero para el contig M5N6L415, que es consistente con PHP al predecir que infecta Archaea. La asignación más similar se observó para el contig M6N2L524, que de acuerdo con RaFHA infecta *Euryarchaeota* del género *Haloarcula* y de acuerdo con PHP infecta *Euryarchaeota* del género *Halorubrum*. Por último, RaFAH fue el único método que predijo correctamente el hospedero del virus *Escherichia* ΦX174, que se utilizó como control positivo para la secuenciación del DNA.

Discusión

¿Cómo predecir el mayor número de hospederos de manera correcta?

Aunque RaFAH alcanzó la mayor precisión, sensibilidad y puntuación F1 en la colección de datos de prueba, sólo dos de sus predicciones sobre el conjunto de datos metagenómicos fueron consistentes con las de los métodos basados en espaciadores CRISPR o en perfiles de oligonucleótidos (PHP), o con el entorno a partir del cual se generaron los metagenomas. Los datos metagenómicos analizados aquí proceden de muestras tomadas dentro de la CCB que, a pesar de ser un oasis desértico con aguas oligotróficas, es conocida por albergar diversos grupos de microorganismos, muchos de los cuales son endémicos y están relacionados con microorganismos marinos (Souza et al., 2006; Souza et al., 2012). Se cree que tal diversidad ha evolucionado como resultado de una larga estabilidad ambiental de un acuífero profundo que recrea las condiciones de un océano antiguo, que nutre a los sistemas acuáticos de la CCB a través de la surgencia del agua subterránea producida por la bolsa magmática en las profundidades de la Sierra de San Marcos y Pinos (Wolaver et al., 2013). Específicamente, el ambiente de donde se extrajeron las muestras es una poza somera caracterizada por alto pH y salinidad conocida como Domos del Arqueano (AD) (Espinosa-Asuar et al., 2022; Medina-Chávez et al., 2023; Cisneros-Martínez et al., 2023). Se ha demostrado que AD alberga una gran diversidad de bacterias en una escala espacial corta (Espinosa-Asuar et al., 2022) y así como una de las comunidades de arqueas más diversas del mundo (Medina-

Chávez et al., 2023). Dicha diversidad incluye bacterias reductoras del sulfato y *Euryarchaeota* halófilas extremas (Madrigal-Trejo et al., 2023). Además, recientemente se ha descrito una comunidad viral muy diversa en la que los haloarqueavirus constituyen una parte esencial (Cisneros-Martínez et al., 2023). Por lo tanto, las predicciones que correspondieron a arqueas halófilas, así como a bacterias halófilas, halotolerantes, alcalófilas, termófilas, oligotróficas, reductoras de sulfato, oxidadoras de azufre o marinas, se consideraron coherentes con el entorno en cuestión.

Aunque CrisprCustomDB fue capaz de discriminar entre posibles hospedadores para el contig C30N1L64 (apoyado además por PHP), el hecho de que el enfoque CRISPR ordinario hiciera más predicciones en el conjunto de datos metagenómicos que CrisprCustomDB probablemente refleje el beneficio de usar una base de datos de espaciadores más extensa (ver Materiales y Métodos) como se explica en el caso del mayor rendimiento de CrisprOpenDB (Dion et al., 2021). Sin embargo, la falta de consistencia de CrisprOpenDB y RaFAH con los otros métodos sugiere que confiar en una base de datos de >11 millones de espaciadores (Dion et al., 2021) o en un clasificador Random Forest basado en el contenido proteico de virus con hospedador conocido (Coutinho et al., 2021), respectivamente, puede ser beneficioso sólo cuando los hospederos o los virus ensamblados ya son conocidos, o están estrechamente relacionados con hospederos o virus representados en las bases de datos correspondientes. Por lo tanto, para conjuntos de datos muy diversos que probablemente tengan una alta proporción de virus nuevos como el que se ha analizado aquí (Cisneros-Martínez et al., 2023), puede ser más apropiado utilizar herramientas basadas en el hospedero, ya sean basadas en la alineación o libres de ella, como CrisprCustomDB o PHP, con bases de datos *ad hoc* construidas con GEMs de arqueas y bacterias del mismo conjunto de datos siempre que sea posible.

Las predicciones sobre el conjunto de datos metagenómicos muestran que métodos fundamentalmente diferentes, como CrisprCustomDB y PHP, pueden complementarse y apoyarse mutuamente. La incorporación de estas herramientas junto con RaFAH, la herramienta de mejor rendimiento en el conjunto de datos de prueba, en un software integrador como iPHoP (Roux et al., 2023), permite abordar el problema de la predicción de hospederos desde diferentes ángulos, aumentando la probabilidad de hacer las predicciones correctas. Además, juzgar las predicciones basándose en la coherencia entre la biología del hospedero predicho (es decir, taxonomía, hábitat, estilo de vida o metabolismo) y el entorno de origen del virus consultado puede proporcionar una validación adicional, principalmente cuando se predicen hospederos de virus nuevos. Sin embargo, hay que tener cuidado con este método de validación. Por ejemplo, para las predicciones con menos coherencia entre métodos y en rangos taxonómicos más altos, existe un mayor riesgo de que la coherencia entre el entorno de origen de los VEMs y la biología de los hospederos predichos sea más bien ambigua o incluso falsa.

La predicción del hospedero es uno de los rasgos más críticos para caracterizar los VEMs, probablemente junto con las relaciones filogenéticas. Queríamos saber quién es el hospedero para conocer mejor la biología del virus recién ensamblado, como de dónde obtiene los recursos para completar su ciclo de replicación, con qué organismos interactúa y con quién podría coevolucionar. Sin embargo, aunque las predicciones de hospederos presentadas aquí nos permiten dar un paso adelante en la caracterización de los virus de AD, aún necesitamos conocer el contexto filogenético, los procesos evolutivos y las adaptaciones funcionales que nos permitirán comprender mejor el origen de la diversidad en este lugar concreto.

Conclusiones

Los resultados presentados aquí indican que RaFAH, un método basado en la alineación dependiente del virus, y PHP, un método sin alineación dependiente del hospedero, son las herramientas de mejor rendimiento para la predicción virus-hospedero. Otros métodos mostraron rendimientos diferentes en función de los criterios de selección del hospedero, los umbrales de puntuación y la base de datos de referencia. Parece que los métodos basados en espaciadores CRISPR se benefician del uso de una base de datos de espaciadores más extensa a la hora de predecir hospederos de virus ya conocidos. Sin embargo, el uso de una base de datos de posibles hospederos más extensa no mejoró el rendimiento de los métodos sin alineamiento dependientes del hospedero, como PHP.

La complementariedad y el apoyo mostrado por CrisprCustomDB y PHP cuando se ejecutaron en VEMs y GEMs del mismo conjunto de datos, sugieren que el uso de tal combinación de herramientas junto con RaFAH puede producir asignaciones de hospederos más fiables en conjuntos de datos metagenómicos altamente diversos, siempre que las predicciones sean consistentes a través de múltiples métodos y la taxonomía, hábitat, estilo de vida o metabolismo del hospedero predicho sea consistente con el ambiente de donde se ensamblaron los virus.

Por último, las predicciones de hospederos sobre los VEMs de los Domos del Arqueano mostraron que los virus que habitan en ese ambiente infectan a arqueas halófilas, así como a una variedad de bacterias que pueden ser halófilas, halotolerantes, alcalófilas, termófilas, oligotróficas, reductoras de sulfatos o relacionadas con el medio marino. Estas predicciones son coherentes con el ambiente particular y con la evolución geológica y biológica de CCB y sus microorganismos.

Artículo requisito: Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat

MICROBIAL GENOMICS

RESEARCH ARTICLE
Cisneros-Martínez *et al., Microbial Genomics* 2023;9:001063
DOI 10.1099/mgen.0.001063

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat

Alejandro Miguel Cisneros-Martínez[1,2], Luis E. Eguiarte[1] and Valeria Souza[1,3,*]

Abstract

In spring 2016, a shallow hypersaline pond (50×25 m) was found in the Cuatro Cienegas Basin (CCB). This pond, known as Archaean Domes (AD) because of its elastic microbial mats that form dome-shaped structures due to the production of reducing gases reminiscent of the Archaean eon, such as methane and hydrogen sulfide, harbour a highly diverse microbial community, rich in halophilic and methanogenic archaea. AD is a seasonally fluctuating hypersaline site, with salinity ranging from low hypersaline (5.3%) during the wet season to high hypersaline (saturation) during the dry season. To characterize the viral community and to test whether it resembles those of other hypersaline sites (whose diversity is conditioned by salinity), or if it is similar to other CCB sites (with which it shares a common geological history), we generated 12 metagenomes from different seasons and depths over a 4 year period and compared them to 35 metagenomes from varied environments. Haloarchaeaviruses were detected, but were never dominant (average of 15.37% of the total viral species), and the viral community structure and diversity were not affected by environmental fluctuations. In fact, unlike other viral communities at hypersaline sites, AD remained more diverse than other environments regardless of season. β-Diversity analyses show that AD is closely related to other CCB sites, although it has a unique viral community that forms a cluster of its own. The similarity of two surface samples to the 30 and 50 cm depth samples, as well as the observed increase in diversity at greater depths, supports the hypothesis that the diversity of CCB has evolved as a result of a long time environmental stability of a deep aquifer that functions as a 'seed bank' of great microbial diversity that is transported to the surface by sporadic groundwater upwelling events.

DATA SUMMARY

Sequence reads are available on the National Centre for Biotechnology Information (NCBI) Sequence Reads Archive. A full list of accession numbers for metagenomes used in this study are available in File S1, available in the online version of this article. Data processing scripts are publicly available on https://github.com/AleCisMar/COMETS. Supplementary materials are available on Figshare: https://doi.org/10.6084/m9.figshare.20958184[1].

INTRODUCTION

Cuatro Cienegas Basin (CCB thereafter) is an endangered oasis in the Chihuahuan desert of Mexico, with N:P ratios in the aquatic environments far from the 16:1 Redfield ratio, ranging from very low phosphorus (157:1) to very low nitrogen (2:1) [2].

Received 08 October 2022; Accepted 15 June 2023; Published 17 July 2023

Author affiliations: [1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, México; [2]Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, Ciudad de México, México; [3]Centro de Estudios del Cuaternario de Fuego-Patagonia y Antártica (CEQUA), Punta Arenas, Chile.
***Correspondence:** Valeria Souza, souza@unam.mx

Keywords: Cuatro Cienegas Basin; hypersaline environment; metagenomics; virome; viral diversity.

Abbreviations: AD, archaean domes; CCB, cuatro cienegas basin; DOM, dissolved organic matter; LCA, least common ancestor; NCBI, National Centre for Biotechnology Information; NMDS, non-metric multidimensional scaling; OC, orange circles; OTU, Operational Taxonomic Unit; SRA, sequence read archive.

Sequence reads are available on the National Centre for Biotechnology Information (NCBI) Sequence Reads Archive. All generated metagenomes are available under the BioProject accession: PRJNA847603.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files.Eight supplementary figures, three supplementary tables and one supplementary file are available with the online version of this article.

Impact Statement

Cuatro Cienegas Basin (CCB) is an endangered oasis in the Chihuahuan Desert of Mexico that, despite its oligotrophic status, is known to harbour great biodiversity, including animals, plants, fungi, and microbes. The antiquity of its sediments, along with the presence of microbial lineages endemic to CCB, adapted to a stoichiometry reminiscent of the Precambrian supereon and related to marine organisms (from which they diverged more than 600 million years ago) and the local abundance of organo-sedimentary structures (such as microbial mats and stromatolites), have positioned CCB as an analogue of the early Earth and as an Astrobiological park. Archaean Domes (AD) is a unique site within CCB that will deepen our understanding of the origin, diversity and dynamics of microbial communities at this remarkable site. The results presented here show that viral communities at hypersaline sites may be subject to local processes that differentiate them from viral communities at other hyperaline sites around the world, and that viral diversity at different CCB sites may be connected through and nurtured by deep aquifer movements.

Despite its oligotrophic status, CCB is known to harbour a great biodiversity, including animals, plants, fungi and microbes and has been suggested to be an analogue of early Earth for different reasons, including the antiquity of its sediments [3] and the local abundance of organosedimentary structures, such as microbial mats and stromatolites, which are indeed conspicuous in the fossil record dating back to the Archaean eon [4]. The idea that CCB is a model of early Earth is reinforced by the presence of endemic microorganisms, adapted to a stoichiometry reminiscent of the late Precambrian supereon [5], related to marine organisms from which they are estimated to have diverged between 770–680 and 202–160 million years ago [6, 7]. Isotopic studies have shown that CCB aquatic systems are largely composed of aquifer groundwater [8] suggesting that the diversity of the CCB has evolved as a result of long-standing environmental stability of a deep aquifer that recreates ancient ocean conditions [7, 8]. As a result, CCB is considered not only a relevant site for the study of early evolutionary and ecological processes – as illustrated by the discovery of new species and the description of adaptations to extreme environments [7] – but also as an Astrobiological park for the identification of biosignatures that can be used in the search for extraterrestrial life [2].

In spring 2016, a small (50×25 m) and shallow alkaline and hypersaline pond was found in CCB (Fig. 1a). This particular site has ellipsoid cisterns filled with orange water, that we named 'orange circles' (OC thereafter) (Fig. 1c, d), that are wet all year and that are particularly rich in clay and silt. Around those OC, in the interphase with sandier sediment, microbial mats form under a salty crust. In that year, after a heavy rain dissolved such salty crust, the microbial mats started to bulge with anoxic gases reminiscent of the Archaean eon, such as methane and hydrogen sulfide [9], forming dome shaped structures (Fig. 1b) and suggesting a deeper connection with the deep aquifer and its deep biosphere. Because of their unique elastic microbial mats and the 'Archaean like atmosphere' that was found inside the domes, we called the site "Archaean Domes" or AD [9, 10]. Given that there are low quantities of Cu metalloenzymes, which is consistent with the low bioavailability of Cu during the Archaean eon [11], and that extreme pH and salinity environments are considered models of ancient Martian ecosystems [12, 13], AD could be considered as a possible model of very ancient Earth communities and a site of astrobiological interest.

AD is a seasonal fluctuating site, with the N:P ratio, pH and salinity ranging from 10:1, 9.5 and low hypersaline (5.3%) during the wet season (Fig. 1d) to 78:1, 5.5 and high hypersaline (saturation) during the dry season (Fig. 1c), respectively [9, 10]. Despite its fluctuating and extreme nature, AD seem to harbour a seasonally stable microbial core community, with some degree of functionally redundant taxa [10] and a high abundance of alkaline and salt resistance genes [14]. AD is also very diverse in microbes, including more than 6000 ASV in ten samples obtained at a scale of 1.5 m [10]. This diversity includes a high abundance of halotolerant bacteria, as well as halophilic and methanogenic archaea [9, 10]. AD is particularly rich and diverse in archaea, which are rare in the rest of CCB [9, 10].

Hypersaline environments are characterized by higher salt concentrations than those of seawater (3–4%) [15] and are usually classified according to their salinity levels; from low salinity (< 10% NaCl), to intermediate salinity (10%–20% NaCl) and high salinity (> 20% NaCl) [16]. In turn, halophilic organisms are classified according to the NaCl concentration that they require for an optimal growth: i) slight halophiles (1–5%); ii) moderate halophiles (5–20%), and iii) extreme halophiles (20–30%) [15]. Some halophilic microorganisms are also high pH-loving organisms or alkaliphiles [17]. Alkaliphiles can thrive in environments with pH >9 [18], however haloalkaliphilic organisms are typically found in saline alkaline lakes, also known as soda lakes. Soda lakes are characterized by highly alkaline water (pH >9) as a result of high carbonate alkalinity (high CO_3^{2-} and HCO_3^- concentrations) coupled with low Ca^{2+} and Mg^{2+} [19, 20].

Similarly to thalassic hypersaline environments with neutral pH, soda lakes microbial community composition is strongly influenced by salinity, where higher salt concentrations result in highly abundant extreme halophilic archaea belonging to the class *Halobacteria* and thus in a less diverse community [21–23]. However, soda lakes tend to be more diverse than neutral pH hypersaline environments, which is likely related to the high availability of CO_2 for primary producers [24] and low Ca^{2+} and Mg^{2+}

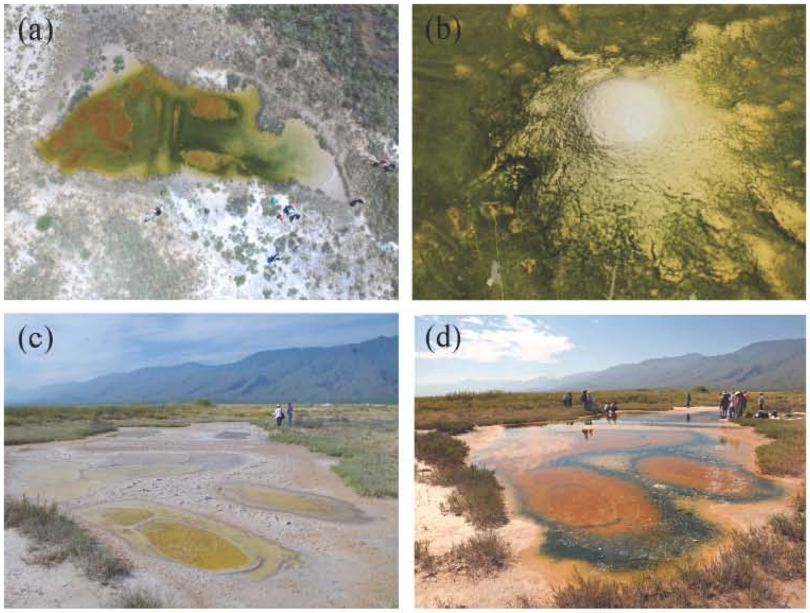


Fig. 1. Shallow hypersaline pond, named Archaean Domes or AD from the Cuatro Cienegas Basin (CCB), Mexico. (a) Aerial view of the small pond (50×25 m). (b) Elastic microbial mat forming a dome shaped structure with a saline crust on top. (c) AD during the dry season (April 2016) (d) AD during the wet season (September 2019). In C and D Orange Circles or OC are visible.

concentrations [23]. Soda lakes are populated by various salinity and pH adapted bacteria and archaea including members of the *Halomonas* group, well represented strains related to *Bacillus alcalophilus* [24] and at higher salt concentrations *Euryarchaeota* of the class *Halobacteria*, the order *Methanosarcinales* [23] and the genera *Natronococcus* and *Natronobacterium* [17, 24] are readily available.

These trends are likely to extend to viruses, where it has been observed that both halophilic archaea and its viruses (haloarchaeaviruses) become more abundant along an increasing salinity gradient, resulting in a decreased microbial diversity [25]. Haloarchaeaviruses belong mainly to the order *Caudovirales* (families *Siphoviridae*, *Myoviridae* and *Podoviridae*) with a smaller proportion belonging to other viral families such as *Sphaerolipoviridae*, *Pleolipoviridae* and *Fuselloviridae* [26].

Viruses, which are known as the most abundant and diverse entities in the world, typically outnumbering bacterial abundance by ten-fold, even in oligotrophic environments [27, 28], are abundant in hypersaline environments (4×10^8 – 2×10^9 VLP ml^{-1}) [29]. Viruses are key players in microbial communities, where they can increase the genotypic diversity by mediating genetic exchange among bacterial strains, via transduction, and can also influence host community diversity by selectively killing the densest and most abundant population of fast-growing strains (i.e. kill the winner and similar evolutionary scenarios) [27, 30]. A similar fraction of bacterial mortality is attributed to protist grazing and viral lysis (approximately 20%). However, viral lysis may have a greater impact in the microbial community, by increasing the flux of biomass into dissolved organic matter (DOM) and thus decreasing biomass transfer to higher trophic levels (viral shunt) [27, 28]. Recycling of DOM can produce significant changes in the nutrient pool usually stimulating bacterial growth [27]. This may have a greater impact in oligotrophic conditions, where biomass consists primarily of microbes, and on hypersaline environments, where protist grazing has been seen to disappear at salinities greater than 20% [31], and therefore haloarchaeaviruses are expected to play the most important role in controlling the halophilic microbial communities.

Different metagenomic studies in CCB have described the viral community within stromatolites from the Río Mezquites river, thrombolites from Pozas Azules II pond [32] and water samples from Churince, La Becerra and Pozas Rojas ponds [33]. These communities are typically dominated by dsDNA bacteriophages from the *Caudovirales* families *Siphoviridae*, *Myoviridae* and *Podoviridae*, followed by ssDNA bacteriophages from the family *Microviridae*, a variety of DNA and RNA viruses from different eukaryotic viral families and virtually no virus infecting archaea [33]. Together, these studies show that the CCB viral communities tend to reflect the diversity patterns of their hosts, displaying a high diversity within and between sites, taxonomic similarity to marine samples and strong signals of endemism [32, 33].

Although viruses are expected to play a significant role in AD microbial dynamics, its viral fraction has not been explored yet. We believe that AD is a microbial community similar to the ones observed in Buck Reef Chert 3.4 Gy fossils in South Africa [34], and that the study of the viruses at AD and their dynamics in time and space will help us to understand the early diversification of life. Given the high salinity of AD, we expected to find a viral community unlike any other within CCB, similar those in other hypersaline environments dominated by haloarchaeaviruses of the order *Caudovirales* [26], and a distribution of diversity in accordance with the reported global patterns of hypersaline viral communities that are driven by salinity levels [25]. Therefore, given the fluctuating conditions of AD, we expected to find clear differences in the community structure between wet and dry seasons. Alternatively, we expected to find a viral community similar to those of other CCB sites due to the influence of groundwater movement produced by the magmatic pouch in the depths of Sierra San Marcos y Pinos, which is known to be a major source of water for aquatic systems in CCB [8].

Here we describe the structure and diversity of the viral fraction within the AD microbial community by analysing twelve metagenomes derived from samples taken from different seasons and depths over a 4 year period. Consistent with the high abundance of archaea in AD, we describe the first occurrence of haloarchaeaviruses in CCB. Although haloarchaeaviruses appear to be an important component of the community at this hypersaline site, the community does not appear to behave as a canonical hypersaline community. More specifically, the community is not dominated by haloarchaeaviruses and its structure is not driven by environmental fluctuations between wet and dry seasons (and thus changes in salinity), which remains highly diverse even at elevated salinity levels (dry season). AD shows the highest viral diversity compared to other sites in CCB and other reported viromes of the world. We found that despite showing some similarities to other hypersaline viral communities (i.e. the presence of haloarchaeaviruses), the viral community in AD is more similar to other CCB viral communities where it shares a common geological history. The similarity in the taxonomic profiles of metagenomes derived from surface samples from 2019 and 2020 with those of metagenomes derived from samples taken at 30 and 50 cm depth (which tend to be more diverse and show higher haloarchaeavirus abundance) suggests that viral diversity may be affected by the sporadic upwelling of groundwater carrying diverse microorganisms from the depths.

METHODS

Sample collection

Samples were collected inside Rancho Pozas Azules (26°49'41.9"N 102°01'23.6"W) which belongs to Pronatura Noreste, in the Cuatro Ciénegas Basin (CCB), in Coahuila, Mexico (Fig. 1), under SEMARNAT scientific permit number SGPA/DGVS/03121/15.

Samples were collected in April 2016, October 2016, February 2017, October 2018, March 2019, September 2019, and October 2020. Samples taken between February and April correspond to the dry season, while the samples taken between September and October correspond to the wet season. For microbial mats, surface samples were collected by means of a sterile scalpel dissection (8 cm^2 / 40 cm^3) and transferred to 50 ml conical tubes. For deeper samples, 30 cm plastic tubes were used as sediment samplers at depths of 30 and 50 cm. Three samples were collected at the shallow ellipsoid orange pools or orange circles (OC): one superficial water sample on a 50 ml conical tube and two more at depths of 30 and 50 cm, as previously described.

In total 12 samples were taken: four microbial mat superficial samples during the dry seasons (M1, M3, M5; D0); three microbial mat superficial samples during the wet seasons (M2, M4, M6); one superficial water sample at OC during a dry season (C0); two microbial mats deep samples during a wet season (D30, D50) and two OC deep samples during a wet season (C30, C50). All samples were stored in liquid nitrogen until processing.

DNA extraction and sequencing

DNA was extracted according to [35] at the Laboratorio de Evolución Molecular y Experimental of the Instituto de Ecología, Universidad Nacional Autónoma de México, in Mexico City. Briefly, the extractions followed a column based protocol with a Fast DNA Spin Kit for Soil (MP Biomedical). Total DNA was sent to CINVESTAV-LANGEBIO, Irapuato, México for shotgun sequencing with Illumina Mi-Seq paired-end 2×300 technology. The number of raw reads produced for every sequencing is shown in Table 1.

Metagenomic data download

Thirty-five metagenomes available in the literature were downloaded from the Sequence Read Archive (SRA) (File S1, February 2022, https://www.ncbi.nlm.nih.gov/sra), and were grouped into seven types. The metagenomes include 11 previously published CCB viromes [33]; three microbialite viromes [32]; eight high salinity hypersaline viromes [36], one intermediate and one low salinity hypersaline virome [37]; six oceanic viromes [38–40]; and five freshwater viromes [37, 41].

Table 1. AD sample and metagenome features. Superficial microbial mat samples taken between April 2016 and September 2019 during dry (M1, M3 and M5) and wet (M2, M4 and M6) seasons, respectively, were used to analyse seasonal variations. In 2020 three microbial mat samples (D0, D30 and D50) and three orange circle samples (C0, C30 and C50) were taken at 0, 30 and 50 cm depth, respectively, to have an approximation on the possible influence of groundwater upwelling on viral diversity. The number of filtered reads varies from 4412620 in M2 to 26799269 in M1

Sample name	Sample type	Sampling date	Season	Raw reads	Filtered reads	Unclassified %	Bacteria %	Archaea %	Eukaryota %	Viruses %
M1	Microbial mat	April 2016	Dry	28859454	26799269	40.70	56.36	1.89	0.69	0.35
M2	Microbial mat	October 2016	Wet	4772053	4412620	41.67	56.09	1.26	0.72	0.25
M3	Microbial mat	February 2017	Dry	8203484	7484431	37.26	61.03	0.76	0.76	0.19
M4	Microbial mat	October 2018	Wet	10030782	9442166	39.23	58.26	1.58	0.69	0.24
M5	Microbial mat	March 2019	Dry	25873990	24402939	38.64	44.12	16.11	0.66	0.47
M6	Microbial mat	September 2019	Wet	20153088	19486258	37.97	44.71	16.21	0.63	0.48
D0	Microbial mat	October 2020	Wet	17148993	15895120	42.81	50.98	4.85	0.68	0.68
D30	Microbial mat	October 2020	Wet	18976795	18350997	52.22	40.78	5.83	0.82	0.35
D50	Microbial mat	October 2020	Wet	16106607	15418160	51.54	43.19	4.16	0.80	0.30
C0	Water	October 2020	Wet	24065589	22124414	47.25	44.34	7.16	0.56	0.69
C30	Sediment	October 2020	Wet	14315374	13901353	53.33	28.84	16.88	0.55	0.40
C50	Sediment	October 2020	Wet	18050094	17604941	55.42	34.74	8.60	0.71	0.54

The corresponding SRA accessions are: SRX3861423 (CH2, water from Churince, CCB), SRX3861426 (CH4, water from Churince, CCB), SRX3861413 (CH5, water from Churince, CCB), SRX3861414 (CH9, water from Churince, CCB), SRX3861415 (CH10, water from Churince, CCB), SRX3861416 (BE, water from La Becerra, CCB), SRX3861417 (PR1, water from Pozas Rojas, CCB), SRX3861418 (PR3, water from Pozas Rojas, CCB), SRX3861424 (PR4, water from Pozas Rojas, CCB), SRX3861425 (PR7, water from Pozas Rojas, CCB), SRX3861421 (PR9, water from Pozas Rojas, CCB), SRX000208 (PA, thrombolite in Pozas Azules II, CCB), SRX000209 (RM, stromatolite in Rio Mesquites, CCB), SRX000221 (Highborne cay, stromatolite in Highborne cay, Bahamas), SRX117679 (2007At1, high salinity water from hypersaline Lake Tyrell, Australia), SRX117680 (2007At2, high salinity water from hypersaline Lake Tyrell, Australia), SRX117681 (2009B, high salinity water from hypersaline Lake Tyrell, Australia), SRX117682 (2010Bt1, high salinity water from hypersaline Lake Tyrell, Australia), SRX117683 (2010Bt2, high salinity water from hypersaline Lake Tyrell, Australia), SRX117684 (2010Bt3, high salinity water from hypersaline Lake Tyrell, Australia), SRX117685 (2010Bt4, high salinity water from hypersaline Lake Tyrell, Australia), SRX117686 (2010A, high salinity water from hypersaline Lake Tyrell, Australia), SRX000217 (Saltern low, low salinity saltern in San Diego Bay, USA), SRX000218 (Saltern med, intermediate salinity saltern in San Diego Bay, USA), SRX000215 (Tabuaeran atoll, seawater from Pacific ocean), SRX000213 (Palmyra atoll, seawater from Pacific ocean), SRX000206 (Kiritimati atoll, seawater from Pacific ocean), SRX000204 (Kingman atoll, seawater from Pacific ocean), SRX000202 (Sargasso sea, seawater from Sargasso sea), SRX008299 (Tampa bay, seawater from Tampa bay, USA), SRX000211 (TP_1105, freshwater from Tilapia pond in Kent SeaTech, USA), SRX000235 (TP_0506, freshwater from Tilapia pond in Kent SeaTech, USA), SRX000236 (PP_0506, freshwater from Prebead pond in Kent SeaTech, USA), ERX007894 (Lake Pavin, freshwater from Lake Pavin, France) and ERX007895 (Lake Bourget, freshwater from Lake Bourget, France).

Metagenomic comparisons

Metagenome comparisons were computed on a custom script (COMETS: COmpare METagenomeS. available in GitHub https://github.com/AleCisMar/COMETS). Briefly, this script takes a metadata table and all the compressed FASTQ files with single- or paired-end raw reads as input. Then, it uses fastp [42] for adapter removal, filtering of low-quality reads (reads with a maximum of 40% of bases with quality <Q15 are qualified) and deduplication of reads with all identical bases. Next, it uses Kaiju [43] to perform taxonomic classifications by comparing every sequencing read against the nr_euk database. Kaiju uses the Burrows-Wheeler transform to search for matches between sequences of translated reads and the database of microbial coding genes, assigning the taxonomic identifier (from NCBI taxonomy) of the longest exact match to each sequencing read. Kaiju assigns the taxon identifier of the least common ancestor (LCA) if matches of the same length are found in multiple taxa [43]. After the taxonomic classification, COMETS produces a count table and a taxonomy table that are used together with the metadata table to build a phyloseq object [44] in R [45]. In the following step it produces rarefaction curves with the myrlin package [46] and performs a normalization by median sequencing depth.

Finally COMETS generates three types of plots with ggplot2 [47] at different taxonomic levels: taxonomic identifiers assigned by Kaiju are considered OTUs which can be filtered according to taxonomic level, from phylum to species. The first kind of plot is a stacked bar for relative abundance of OTUs that represent at least 1% of the reads in at least one sample; the second type is a dot plot for alpha diversity (Shannon) and; the last is a non-metric multidimensional scaling (NMDS) scatter plot to represent beta diversity (Bray-Curtis dissimilarity). When taxonomically filtered, OTUs without the corresponding taxonomic assignment (NA) were excluded from diversity calculations and plot generation.

To explore the similarities between viral communities at AD and other viral communities in CCB and the rest of the world, we computed Bray-Curtis dissimilarity measures from the normalized count table at different taxonomic levels to build UPGMA trees with the NEIGHBOR programme from the PHYLYP package, using input order and without subreplicates [48].

Statistical analyses

For initial statistical analyses, we assumed a normal distribution and performed ANOVA and one-tailed t-Student tests with alpha=0.05 to contrast different groups of samples, i.e. wet (M2, M4 and M6) and dry (M1, M3 and M5) seasons; early surface (M1, M2, M3 and M4), late surface (M5, M6, D0 and C0) and deep samples (D30, D50, C30 and C50) and; surface (M1, M2, M3, M4, D0 and C0) and deep samples (M5, M6, D30, D50, C30 and C50).

For a more robust analysis of the differential abundance between sample groups, we followed the IDEAmex pipeline originally developed for differential expression analysis [49] with the package edgeR [50]. Normalization was performed with the Trimmed Mean of M-values or TMM method, and differentially abundant OTUs were considered with log fold change=1.5 and *p*-value=0.01.

RESULTS AND DISCUSSION

Description of AD microbial community

We analysed the relative abundance of reads assigned to the three domains of life and viruses to explore the general taxonomic structure of each sampled community in AD. As shown in Table 1, a total of 12 samples were analysed: three microbial mat superficial samples during the dry season (M1, M3 and M5) and three microbial mat superficial samples during the wet season (M2, M4 and M6), to evaluate the effect of environmental fluctuations, and three microbial mat samples (D0, D30 and D50) and three OC samples (C0, C30 and C50) at different depths (0, 30 and 50 cm) (see Methods for details), to indirectly evaluate the effect of the deep aquifer on diversity.

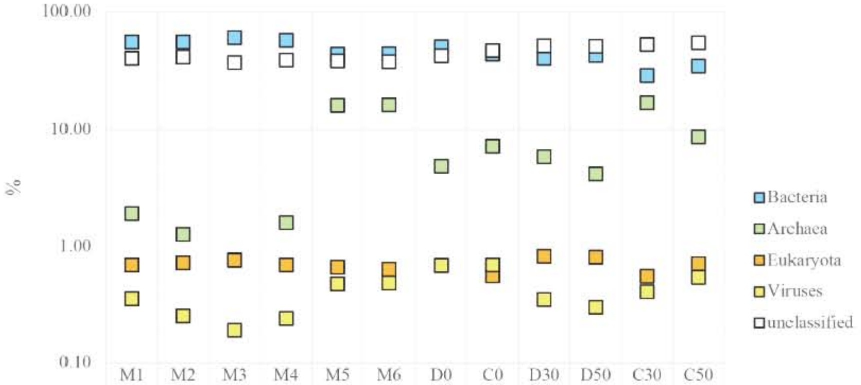
The number of filtered reads was on average 16276889, ranging from 4412620 in M2 to 26799269 in M1. On average, most of the reads are assigned to Bacteria (46.95%), followed by unclassified reads (44.84%), Archaea (7.11%), Eukaryota (0.69%) and Viruses (0.41%). However, these abundances vary over a wide range among the different samples. For example, Bacteria ranged from 61.03% in the surface dry season sample M3, to 28.84% in the 30 cm deep OC sample C30, while Archaea varied from 0.76–16.88% in the same samples (Table 1).

Comparing our samples with the literature, the highest Archaea abundance found in C30 is more similar to the Archaea abundance (16%) reported in Acos, a thalassohaline hypersaline system with intermediate salinity (19%) in Peru, than the abundance described in Maras, another Peruvian site, which can reach more than 30% NaCl, where Archaea can reach up to 80–86% relative abundance [22].

To assess the factors these domain differences can be attributed, we compared the variations in relative abundance among samples grouped by season or depth. We could not find significant differences between dry (M1, M3 and M5) and wet (M2, M4 and M6) seasons samples. However, there were significant changes between some sample groups. For instance, surface samples from 2019 and 2020 (M5, M6, D0 and C0) are distinguished from surface samples from 2016 to 2018 (M1, M2, M3 and M4) by a lower proportion of reads assigned to Bacteria (t-Student, t=5.93, *P*=0.0041), coupled with a greater relative abundance of reads assigned to Archaea (*P*=0.0312) and Viruses (*P*=0.0098) (Fig. 2). Interestingly, surface samples from 2019 onwards only differ significantly from the depth samples (D30, D50, C30 and C50) by a lower percentage of unclassified reads (*P*=0.0077). However, until more samples are obtained for comparison, interpretations should be taken with reservation.

On average, the microbial community was dominated by bacteria of the phylum *Proteobacteria* (44.32%), followed by *Bacteroidetes* (10.66%), *Cyanobacteria* (10.49%) and *Chloroflexi* (9.14%). *Cyanobacteria* reached their highest abundance in the shallow samples reaching dominance in M3 (45.08%), but suffering a large decrease in M5 and M6 (on average 1.96%) and in deep samples (on average 0.18%). *Chloroflexi* maintained an abundance of less than 10% in most samples but appeared to become more abundant with depth, reaching an average abundance of 34.26% at D30 and D50. It is also important to note that *Euryarcaheota* had an abundance of less than 0.1% in surface samples from 2016 to 2018 but reached an average abundance of 14.26% at M5 and M6, after which their abundance never again became as low (Fig. S1).

The viral community was, on average, dominated by reads assigned to the order *Caudovirales*, namely to the families *Siphoviridae* (53.35%), *Myoviridae* (31.48%) and *Podoviridae* (11.92%). These viral families are also ubiquitous in hypersaline sites [25] and other extreme environments, except for deep sea sediments and cold environments [51]. Viruses of the order *Caudovirales* – also known as head-tailed viruses – are the most abundant type of viruses on Earth [52], and are non-enveloped viruses with



	M1	M2	M3	M4	M5	M6	D0	C0	D30	D50	C30	C50
Bacteria	56.36	56.09	61.03	58.26	44.12	44.71	50.98	44.34	40.78	43.19	28.84	34.74
Archaea	1.89	1.26	0.76	1.58	16.11	16.21	4.85	7.16	5.83	4.16	16.88	8.60
Eukaryota	0.69	0.72	0.76	0.69	0.66	0.63	0.68	0.56	0.82	0.80	0.55	0.71
Viruses	0.35	0.25	0.19	0.24	0.47	0.48	0.68	0.69	0.35	0.30	0.40	0.54
unclassified	40.70	41.67	37.26	39.23	38.64	37.97	42.81	47.25	52.22	51.54	53.33	55.42

Fig. 2. Percentage of AD reads assigned to Bacteria, Archaea, Viruses or unclassified. Surface samples from 2016 to 2018 (m1-m4) have a larger abundance of reads assigned to Bacteria and a lower abundance of reads assigned to Archaea and Viruses. Metagenomes sampled at depths of 30 and 50 cm (d30, d50, c30, c50) show the largest proportion of unclassified reads. Absolut values corresponding to 100% are presented in Table 1.

ichosahedral heads or capsids attached to a hollow flexible tail. They account for 96% of viruses infecting bacteria [53] and the majority of viral isolates infecting halophilic or methanogenic euryarchaea, accounting for almost half of all archaeal viruses studied [54]. These proportions are somewhat similar to other CCB viromes [33], albeit with a lower proportion of viruses belonging to the family *Microviridae* (0.51%) and a subtle presence of *Herelleviridae* (1.15%) and *Marseilleviridae* (1.56%) (Fig. S2A).

At genus level, AD had a high abundance of *Donellivirus* (36.29%), which includes *Bacillus phage G*, initially isolated from its host *Bacillus megaterium* [55], followed by *Ahduovirus* (17.74%), which includes *Burkholderia phage vB_BceS_AH2*, originally isolated from plant-associated soil samples [56], *Barbavirus* (12%), which includes various *Rheinheimera* sp. phages isolated from Baltic Sea samples [57], *Emdodecavirus* (11.12%), including *Sinorhizobium phage phiM12* that infects *Sinorhizobium meliloti* [58], *Shapirovirus* (7.03%), which includes various *Caulobacter crescentus* phages, some of which have been isolated from superficial freshwater samples [59], and *Bellamyvirus* (6.52%), comprising the *Synechococcus phage Bellamy*. These are all head-tailed viruses which, as seen in Fig. S2B, can also be found in hypersaline viromes (*Donellivirus*, *Barbavirus* and *Bellamyvirus*), in other CCB sites (*Barbavirus*, *Embdodecavirus* and *Shapirovirus*) and ocean viromes (*Bellamyvirus*), except for *Ahduoviruses* which are rare or absent in other viromes. *Bacillus megatrium* and *Sinorhizobium meliloti* are present in the AD metagenomes as are numerous reads assigned to members of the genera *Burkholderia*, *Rheinheimera*, *Caulobacter* and *Synechococcus*, suggesting that the aforementioned viruses may indeed be present and infecting their corresponding hosts at this site.

As seen in Fig. 2, late surface samples (M5, M6, D0 and C0) had a larger abundance of reads assigned to Archaea and Viruses when compared to early surface samples. This trend can be further explored for viral reads with successful taxonomy assignment at species level. For instance, from samples taken during 2019 (M5 and M6) onwards, an increase in reads assigned to haloarchaeaviruses also found at other hypersaline sites could be observed (Fig. 3a).

For surface samples from 2019 onwards, there was a considerable decrease in reads assigned to *Microviridae* sp., *Circoviridae* sp., *Microvirus* sp., Prokaryotic dsDNA virus sp., uncultured marine phage and *Synechococcus* phage S-SCSM1, which were also less abundant at the samples of 30 and 50 cm depth, coupled with a considerable increase in reads assigned to viruses that infect halophilic Archaea and some environmental halophages with no known host (eHP-6, eHP-15, eHP-20, eHP-28, eHP-31 and eHP-34) [53, 60], which are also more abundant in the 30 and 50 cm depth samples. In turn, 30 and 50 cm depth samples had a greater abundance of some reads assigned to viruses that infect halophilic Archaea compared to both surface samples from 2016

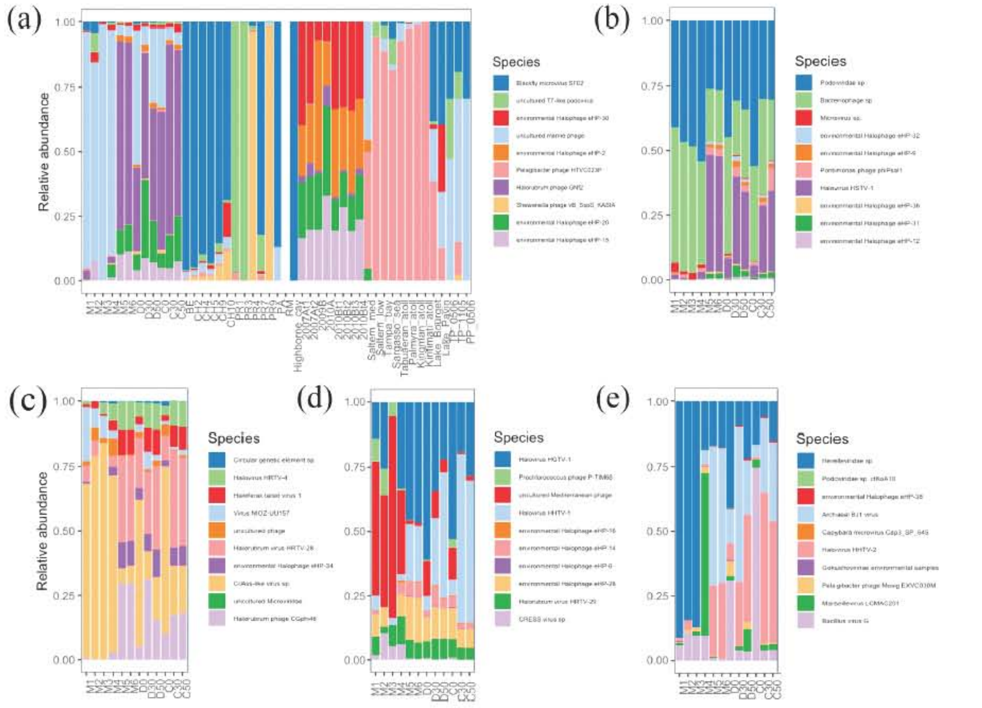


Fig. 3. Relative abundance of reads assigned to viral species. For clarity, 91 species were sorted in descending order of mean relative abundance obtained from the 47 metagenomes, and are presented in separate graphs in groups of ten. (a) Fifth group of most abundant species for 12 AD metagenomes and 35 metagenomes from other environments (see Table S1). From M5 onwards, a high abundance of reads assigned to haloarchaeaviruses, which are also found in other hypersaline environments, can be observed. (b) Second group of most abundant species for 12 AD metagenomes. (c) Third group of most abundant species for 12 AD metagenomes. (d) Fourth group of most abundant species for 12 AD metagenomes. (e) Seventh group of most abundant species for 12 AD metagenomes. (b–e) show that M5 and M6 have abundances similar to those of 30 and 50 cm depth samples. Absolute values corresponding to 100% range from 44067 in 2009B to 97972 in BE, with an average of 76045. For AD metagenomes the variation is from 62939 in D50 to 73174 in M1, with an average of 69740.

to 2018 and from 2019 onwards (Table S1). Although the number of samples in each sampling time group was low (n=3), the magnitude of the changes strongly suggests that the differences are significant.

Interestingly, among surface samples from 2019 onwards, M5 and M6 showed haloarchaeavirus abundances more similar to those at 30 and 50 cm deep samples (Fig. 3b–e). To further explore this trend, we performed a hierarchical clustering based on the relative abundance of reads assigned to viral species with at least 1% average abundance across all AD samples. This analysis suggested that the AD samples could be divided into two groups: the surface samples, including M1-M4, D0 and C0; and the deep samples (*sensu lato*), which included the 30 and 50 cm deep samples along with M5 and M6 (Fig. 4). A comparison between AD deep *sensu lato* and surface samples showed significant differences (t-Student, $P<0.05$) in 45 out of 86 OTUs with species assignment in AD viromes (Table S2). These are 12 OTUs more abundant in surface viromes, including cyanophages, and 33 OTUs more abundant in deep viromes *sensu lato*, including all OTUs assigned to haloviruses.

As seen in Table 2, surface samples from 2016 to 2018 are characterized by a low haloarchaeavirus abundance (< 2%), while surface samples from 2019 to 2020 and deep samples have haloarchaeavirus abundance greater than 3%. However, M5 and M6 reach the haloarchaeavirus abundance that can be found in the deep samples (> 20%). On average, haloarchaeaviruses reach 15.37% relative abundance, which is higher than what is found in low hypersaline sites (3%) but lower than what can be found in high hypersaline sites (up to 76%) [25].

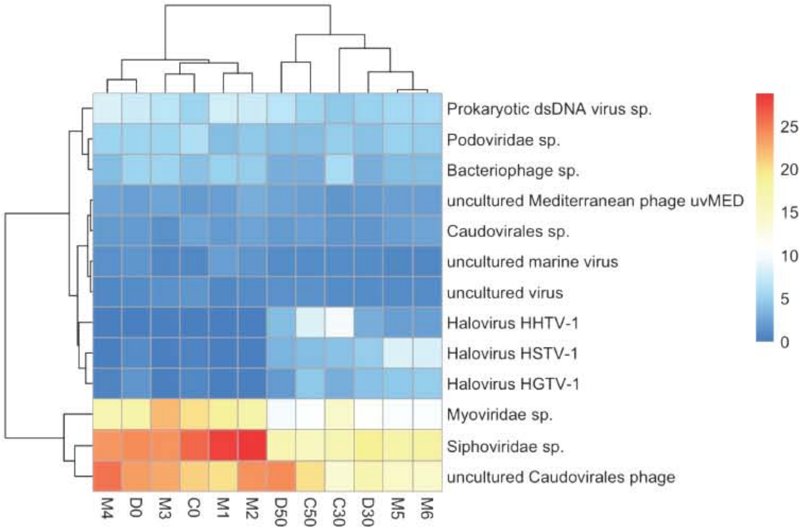


Fig. 4. Hierarchical clustering based on relative abundance of reads assigned to viral species with at least 1% average abundance across all AD metagenomes. Most surface samples cluster together, including D0 and C0 which are late surface samples (2020), except for M5 and M6 (2019) which cluster with 30 and 50 cm depth samples. Absolute values corresponding to 100% range from 50 897 in D50 to 66 360 in M1, with an average of 60 028.

Since these taxonomic analyses were based on the relative abundance of filtered abundant OTUs with successful taxonomic assignments (see Methods), we also performed a more exhaustive analysis, including all OTUs at AD to evaluate differential abundance between surface and deep *sensu lato* samples (an equivalent analysis was performed between dry and wet samples, showing no significant differentiation; analysis not shown). Briefly, we used the raw counts table and followed IDEAmex pipeline

Table 2. AD samples with their sampling year, sampling depth, the number of reads assigned to viral species, number of reads assigned to haloarchaeavirus and the relative abundance of haloarchaeavirus. The relative abundance of haloarchaeaviruses ranges from 0.17% in M3 to 32.20% in C30

Sample	Year	Depth (cm)	Reads assigned to viral species	Reads assigned to haloarchaeaviruses	Haloarchaeavirus abundance (%)
M1	2016	Surface	73174	652	0.89
M2	2016	Surface	68923	321	0.47
M3	2017	Surface	70741	123	0.17
M4	2018	Surface	67637	1200	1.77
M5	2019	Surface	71337	21156	29.66
M6	2019	Surface	71146	20945	29.44
D0	2020	Surface	68449	3752	5.48
D30	2020	30	69594	21088	30.30
D50	2020	50	62939	12913	20.52
C0	2020	Surface	71094	2343	3.30
C30	2020	30	70555	22719	32.20
C50	2020	50	71293	21544	30.22

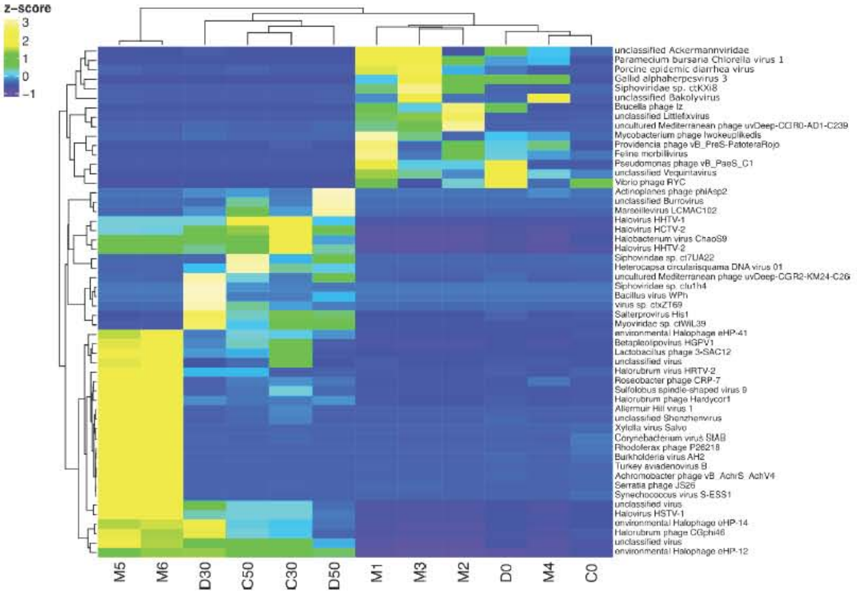


Fig. 5. Heatmap representing differentially abundant viral OTUs of AD at CCB, using all OTUs to evaluate differential abundance between surface and deep samples (log fold change=1.5 and *p*-value=0.01) between deep (*sensu lato*) and surface samples.

for differential expression analysis [49] with the package edgeR [50]. There were 54 out of 3228 differentially abundant OTUs (log fold change=1.5 and *p*-value=0.01). Of these, 39 were more abundant in deep samples, and 15 were more abundant in surface samples (Fig. 5). Of the abundant OTUs in deeper samples, 14 were assigned to archaeal viruses, of which 13 belonged to haloarchaeaviruses, and one to an acidophilic and hypertherophilic archaea virus. There were no OTUs assigned to haloviruses or viruses of archaea significantly more abundant in surface samples. Within OTUs with non-significant differential abundances, there were 98 haloviruses of which 95 showed a tendency to a greater abundance in deep samples. Interestingly, although M5 and M6 viromes were clearly differentiated from other surface viromes and shared some characteristics with deep viromes, there were some differentially abundant OTUs which were in low abundance in M5 and M6 (as in surface viromes) and others which were only abundant in M5 and M6 but not in depth or surface viromes.

Consistent with Table 2, 2020 surface viromes (D0 and C0) showed a less drastic decrease in OTUs assigned to haloviruses compared to other surface viromes. In fact, in a MDS of normalized counts M5 and M6 could be somewhat differentiated from both surface and depth viromes, and D0 and C0, although clustered with other surface viromes, were slightly closer to deep viromes (Fig. S3).

Viral diversity in AD and comparison with other viromes

Within AD, Shannon's diversity index was greater for 30 and 50 cm depth viromes in comparison to both early (t-Student, t=4.44, *P*=0.0044) and late (t-Student, t=3.04, *P*=0.0228) surface viromes (Fig. 6). Estimated Richness (Chao1) was significificatively higher (t-Student, t=2.68, *P*=0.0368, Fig. S4A) for late 2019–2020 viromes of AD in comparison to early 2016–2018 AD viromes, while Evenness (Simpson) was greater (t-Student, t=9.9, *P*=6.12e-05, Fig. S4B) in 30 and 50 cm depth viromes compared to early 2016–2018 viromes.

Since, as shown above, the 2019 surface samples (M5 and M6) showed higher similarities with the 30 and 50 cm depth samples, we grouped the 2019 surface samples with the 30 and 50 cm depth samples (deep samples, *sensu lato*) and compared them against the rest of the surface samples. Both the Simpson and Shannon indices had higher values for deep viromes *sensu lato* in comparison to surface viromes (*P*=2.12e-07 and *P*=0.0014, respectively). There are no differences between dry and wet seasons.

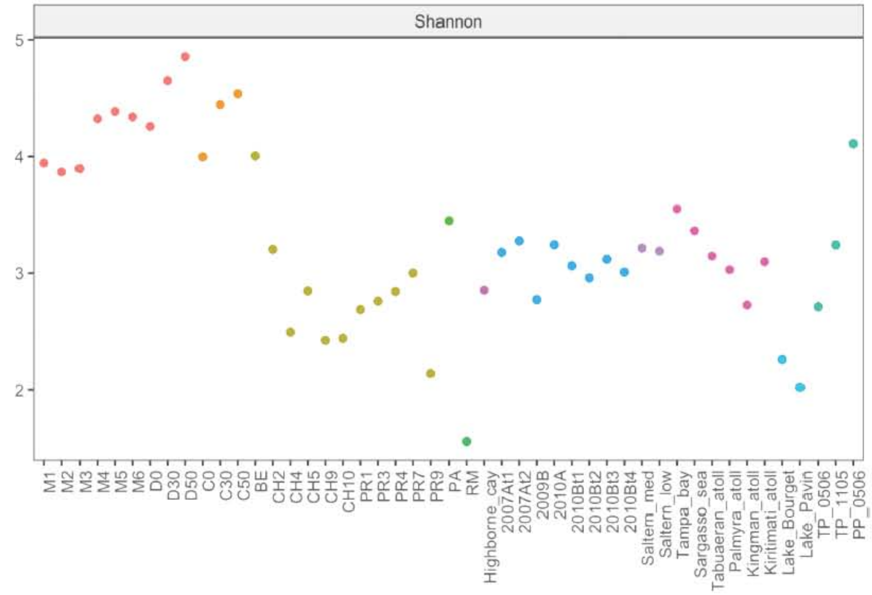


Fig. 6. OTU level alpha diversity index (Shannon) for 12 viral metagenomes from AD, at CCB, and 35 metagenomes from other environments (see File S1). AD viromes are represented by red points. Other CCB viromes (PR and CH) are represented by olive green points. Sea viromes are represented by pink points. High hypersaline viromes are represented by blue points.

Alpha diversity estimates showed that AD harbour the most diverse viral community among all other available viromes, including CCB, hypersaline, sea and freshwater samples (Shannon Fig. 6; Chao1 Fig. S4A and Simpson Fig. S4B, all at OTU level). More specifically, AD show significantly higher Shannon index values than those of Churince (t-Student, t=9.12, *P*=3.92e-05), Pozas Rojas (t-Student, t=9.33, *P*=3.38e-05), high hypersaline (t-Student, t=11.25, *P*=2.67e-09) and ocean (t-Student, t=7.75, *P*=8.79e-06) viromes.

The Bray-Curtis clustering showed that AD viromes were more similar to other CCB viromes than to hypersaline viromes, and that AD formed a cluster of their own within which samples were grouped by depth rather than seasons (Fig. 7). Consistent with previous groupings, M5 and M6 were clustered with 30 and 50 cm deep viromes, while 2020 surface viromes (D0 and C0) were clustered with the early AD surface viromes.

A NMDS analysis based on Bray-Curtis dissimilarities (Fig. S5) between the 47 viromes, supported the analysis described above, showing that all AD viromes shared a unique viral community despite seasonal differences in pH and salinity and that neither AD viromes derived from saturation conditions during the dry season cluster with other highly hypersaline viromes, nor do AD viromes derived from lower (5%) hypersaline conditions during the wet season clustered with other low or intermediate hypersaline viromes.

For a more in-depth analysis, we transformed the OTU level Bray-Curtis dissimilarity matrix into a similarity matrix (1 – dissimilarity) and visualized the similarity distribution on networks built with the igraph library [61], as shown in Fig. 8. Within the 15.9% (above one standard deviation) of the strongest similarities, four different clusters could be found: Cluster 1, including all AD viromes; Cluster 2, including all high hypersaline viromes; Cluster 3, with all ocean viromes plus the low salinity hypersaline virome and; Cluster 4, with the other CCB viromes (all from Churince and Pozas Rojas) along with four freshwater viromes. In Cluster 4 Pozas Rojas viromes are also connected to the intermediate salinity hypersaline virome and one Churince virome is also connected to Pozas Azules and La Becerra viromes. Within the 25% (above third quartile) of the strongest similarities, high-salinity hypersaline viromes remained as an isolated cluster, while AD viromes showed connections with Cluster 4, namely between AD surface and Churince viromes (Fig. 8). At this point, Cluster 3 was also connected to Cluster 4 through Pozas Rojas and intermediate salinity hypersaline viromes.

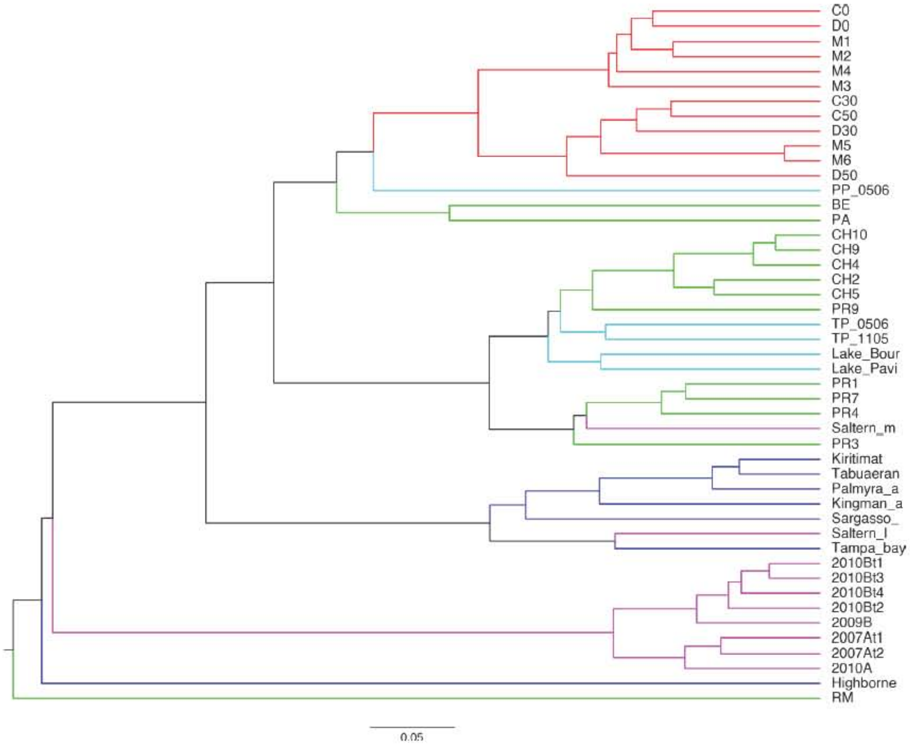


Fig. 7. UPGMA tree computed from normalized species-level Bray-Curtis dissimilarity matrix. This matrix is based on the relative abundance of filtered abundant OTUs with successful taxonomic assignments. Branches leading to AD viromes are highlighted in red. Branches leading to other CCB viromes are highlighted in green. Freshwater, ocean and hypersaline viromes are highlighted by cyan, blue and magenta branches, respectively. This tree shows that AD harbours a unique viral community more similar to those from other CCB sites than to those from other hypersaline sites.

When the top 50% (above second quartile) similarities were considered (Figure not shown), Cluster 1 and Cluster 4 appeared as a tightly packed cluster with a weaker association to Cluster 3 (via Pozas Rojas viromes) and no association to Cluster 2. If the top 75% (above first quartile) similarities were considered, Cluster 1, Cluster 4 and Cluster 3 merged as a single cluster with a weaker similarity to Cluster 2 mainly due to AD deep viromes (Fig. S6).

Overall, CCB viromes (Churince and Pozas Rojas) were similar to freshwater viromes, especially those from Churince (average Bray-Curtis=0.3423). Also, AD viromes were more closely related to Churince (average Bray-Curtis=0.6378), Pozas Rojas (average Bray-Curtis=0.6607) and freshwater viromes (average Bray-Curtis=0.6863), than to oceanic (average Bray-Curtis=0.7440), and high salinity hypersaline viromes (average Bray-Curtis=0.8788). Interestingly, Pozas Rojas viromes were more closely related to oceanic viromes (average Bray-Curtis=0.6285) than to those from AD. Other studies in CCB have described a high β-diversity for viral [32, 33] and microbial [62] communities. However, those studies either lacked appropriate methods for comparison or did not include enough metagenomes from other environments to establish a scale that would put diversity differences between different CCB sites into context. More recent studies have recognized that CCB viral communities had similarities to some marine communities, as previously described [32], and some freshwater communities [33], and a close relationship between Churince and Pozas Rojas microbial communities, with some Pozas Rojas samples closely related to the epipelagic zone of the Mediterranean Sea [63]. Here we showed that, although highly diverse, CCB environments may actually harbour a group of somewhat related microbial communities.

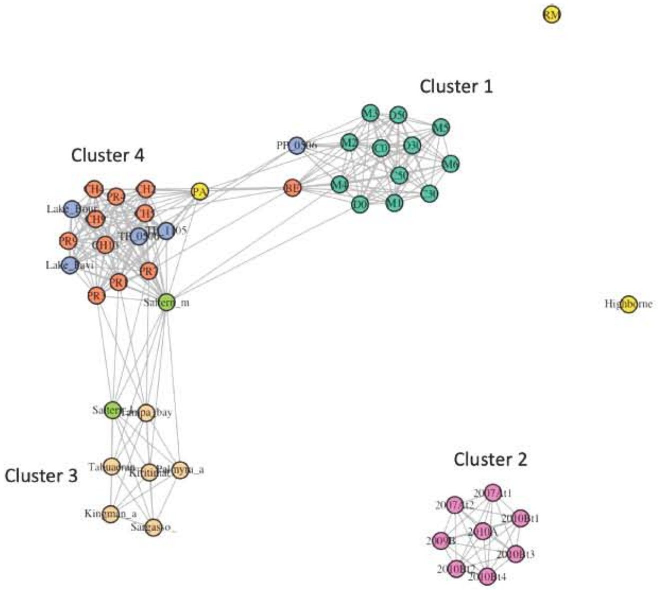


Fig. 8. OTU level Bray-Curtis similarity network showing 25% (above third quartile) of the strongest similarities. AD viromes are represented by jade green circles within Cluster 1. Other CCB viromes (PR and CH) are represented by orange circles within Cluster 4. Ocean viromes are represented by beige circles within Cluster 3. High hypersaline viromes are represented by pink circles within Cluster 2.

Another interesting trend is that AD viromes are the least dissimilar from high salinity hypersaline viromes. More strikingly, viromes from deeper samples (including M5 and M6) were significatively less dissimilar (P=3.6e-32) from high salinity hypersaline viromes (average Bray-Curtis=0.8354, n=48) than surface viromes (average Bray-Curtis=0.9223, n=48). Deep viromes *sensu stricto* were also significatively less dissimilar (P=6.26e-05) from high salinity hypersaline viromes (average Bray-Curtis=0.8348, n=32) than late surface viromes (average Bray-Curtis=0.8692, n=32) which in turn were less dissimilar (P=4.17e-11) than early surface viromes (average Bray-Curtis=0.9326, n=32). This trend highlights the fact that, although not closely related to viral communities from other high hypersaline sites, AD is a hypersaline site harbouring a considerable abundance of halarchaeaviruses that become more abundant and diverse at greater depths.

Given that AD harbours microbial mats associated with elevated salinity and pH, it is likely that the viral community resembles that of other hypersaline microbial mats or that of soda lakes. To test this hypothesis, we downloaded five metagenomes from hypersaline microbial mats and soda lakes, respectively (Table S3), and ran the COMETS pipeline (see Methods) to conduct an alpha and beta diversity analysis in perspective with all previous metagenomes. The OTU level Bray-Curtis similarity network showing 25% (above third quartile) of the strongest similarities (Fig. S7) had an overall structure very similar to that of Fig. 8, except that high hypersaline viromes were now connected to the rest of the viromes through soda lakes viromes. Two soda lake viromes appear closely related to viromes from AD (HC26S and Wadi El-Natrun). However it is four hypersaline microbial mat viromes from other sites (Great Salt Lake, Tristomo elos1, Tristomo elos7 and Tristomo elos12) that appear to group within the AD cluster. Interestingly, these two viromes from soda lakes and four from hypersaline microbial mats, showed alpha diversity indices as high as those from AD (Fig. S8).

Interestingly, the high diversity of soda lakes has inspired the soda ocean hypothesis [24], according to which conditions on the early Earth would have allowed the formation of an alkaline ocean that would have favoured some of the reactions essential for the formation of life and the proliferation of stromatolite-forming organisms [19, 64].

Where does AD viral diversity come from?

The Grinnellian niche concept states that community structure is driven by environmental variables [65, 66]. Hypersaline viral communities have been shown to follow global patterns such that their structure and diversity are driven by changes in salinity levels [25]. Therefore, given the high salinity and the fluctuating conditions of the analysed samples from the Archaean Domes (AD) in the Cuatro Cienegas Basin (CCB), we expected to find a viral community whose structure and diversity, and its response to changes in salinity, would resemble those from other hypersaline sites. We found a community dominated by viruses belonging to the order *Caudovirales* (Fig. S2A), as would be expected from other hypersaline environments [25] and also from other environments from temperate to extreme [51], and a considerable abundance of reads assigned to haloarchaeaviruses (Table 2). Neither the community structure (Fig. 3) nor the alpha diversity (Fig. 6) in AD viral community are driven by environmental fluctuations. For instance, neither samples from the wet season group with low or intermediate salinity hypersaline viromes, nor samples from dry season group with other high hypersaline viromes (Fig. S5). Instead, AD viromes form a cluster of their own, within which the subgroups are sorted by depth rather than season (Fig. 4, Fig. 7). Also, in contradiction to what has been reported in viral communities from other hypersaline sites (regarding the decrease in diversity when the abundance of haloarchaeaviruses is high) [25], in AD samples – in which there is a higher abundance of haloarchaeavirus – higher viral diversity was also observed (Fig. 6). A similar trend with increased microbial diversity at higher salinity, pH and depth has been reported in Ethiopian soda lakes [67].

If environmental variables are not the main community drivers, it is possible that the community dynamics fit an Eltonian niche concept, which states that community structure is driven by interactions [63, 65]. For instance, considering that naked haloarchaeaviruses are only indirectly affected by changes in salinity [53], that AD viral community is dominated by viruses belonging to the order *Caudovirales*, which are naked viruses, and that a highly diverse and seasonally stable core microbial community has been recently described in AD [14], one could argue that AD viral community will tend to remain stable as long as the host community remains the same. This may be related to the so called 'insurance hypothesis', which predicts that highly diverse ecosystems remain functionally stable in changing environments [66]. In addition, organisms inhabiting AD should probably be poikilotrophic, i.e. poly-extremophiles adapted to an environment subject to extreme and sporadic physicochemical changes [68]. Such may be the case for soda lakes inhabited by microorganisms adapted to both elevated pH and salinity. For example, the highly diverse microbial communities of the Kulunda steppe soda lakes, where it has been argued that environmental fluctuations (salinity) promote the maintenance of high diversity [23].

One possible evidence of virus-host interactions in AD is the relatively high abundance of reads assigned to viruses infecting Archaea, which is consistent with the high abundance and diversity of Archaea reported in previous studies [9, 10]. Also, the increase in Archaea abundance to 16% since 2019 may be associated with an increase in virus abundance (Fig. 2), which in turn may contribute to the high diversity and stability of the microbial community via 'kill the winner' interactions [27, 30]. However, further analyses are needed to test the extent and relevance of virus-host interactions in this site.

Most CCB environments have a low carbonate alkalinity [69] and high Mg^{2+} and Ca^{2+} [69–71] which, despite being athalassic, is very similar to seawater ionic composition [70] where Mg^{2+} and Ca^{2+} concentrations are much higher than that of carbonates [20]. Given that AD viral community is closely related to that of other CCB environments (Fig. 7) we could expect the ionic composition to be similar to that of seawater. However, carbonate and bicarbonate measurements on 2019 AD samples resulted in a higher total alkalinity ($TA=2[CO_3^{2-}] + [HCO_3^-]$) (from 11.08 mmol l^{-1} during wet season to 32.75 mmol l^{-1} during dry season) in comparison to seawater (2.33 mmol l^{-1}) [20] and other CCB sites (from ~0.5 mmol l^{-1} to ~6 mmol l^{-1}) [70], but not as high as in Lake Van (~150 mmol l^{-1}), which is the world's largest soda lake [20]. Such alkalinity may be enough to speculate that AD is a soda lake, however this possibility cannot be confirmed until a full anionic/cationic analysis including Mg^{2+} and Ca^{2+} concentrations is made, in order to test the soda lake criterion ($TA>2[Mg^{2+}] + 2[Ca^{2+}]$) [20].

High salinity and pH (up to 9.5 during the wet season), as well as similar community diversity and composition to that of a high hypersaline sediment sample from Hutong Qagan soda lake (Inner Mongolia) (Fig. S7), add to the possibility that AD is a soda lake. However, soda lakes are considered the most stable high pH environments on Earth [19, 24], due to the buffering effect against strong pH variations conferred by high alkalinity [72], which contrasts with the drop in pH to 5.5 reported in AD during the dry season [9], when pH would be expected to increase if it were a true soda lake [20]. In these alkalinity-limited environments, the elevated pH may be the result of net CO_2 removal by photosynthesis during the day. During the night, when there is no photosynthesis, CO_2 is returned to the water through respiration and pH decreases again [72, 73]. Since all samples were taken during the day, we cannot know if this is the case for AD, however it could be a strong possibility since the highest pH is observed during the wet season when photosynthetic cyanobacteria proliferate. Another process that could explain the relatively higher alkalinity and pH at AD compared to other CCB sites is sulphate reduction, which is a proton-consuming process carried out by sulphate-reducing bacteria. Briefly, as sulphate reduction occurs, cyanobacterial mats are degraded and organic matter oxidizes, which results in chlorophyll Mg^{2+} solubilization and bicarbonate production, respectively [73–75]. Both, photosynthetic and sulphate-reducing metabolisms have been detected in AD [14].

The results presented here are in better agreement with the alternative hypothesis that the AD viral community will be more similar to that of other CCB sites due to their shared geological history and deep aquifer. Briefly, after the Pangea breakup, all northern Mexico was covered by a shallow sea that began to regress in the late Cretaceous due to the Laramide Orogeny, completing its regression and the isolation of the CCB from the Gulf of Mexico with the uplift of the Sierra Madre Oriental in the early Eocene [2, 3, 6]. In addition, isotopic studies have shown that deep aquifer groundwater is an important source for CCB aquatic systems

[8], suggesting that the deep aquifer has preserved the conditions of an ancient ocean and has maintained ancient microbial lineages isolated from their marine relatives for millions of years, which has been long enough to allow for such a great microbial diversity to emerge [7, 8].

Although CCB microbial and viral communities have been shown to have high α- and *β*-diversity indices [33, 62], the clustering analyses of the viral communities presented here (Fig. 7) suggest that CCB microbial communities represent a set of related communities. In addition, the fact that AD viral community forms a cluster of its own, within which the subgroups are sorted by depth rather than season (Figs 4 and 7), that diversity increases at greater depths (Fig. 6), and that late surface samples (2019–2020) present traits akin to deep samples (Figs 3 and 5), suggest that the deep aquifer beneath AD harbours a highly diverse microbial community that is sporadically transported to the surface during water upwelling events (maybe moved by the magmatic pouch in the depths of Sierra San Marcos y Pinos [8]). Finally, given that deep viromes show the highest percentage of unclassified reads (Fig. 2), it is likely that the microbial community in the deep aquifer is largely constituted by still unknown microorganisms.

CONCLUSIONS

Overall, these results show that AD harbour a highly diverse and unique viral community rich in haloarchaeaviruses. Although the presence of haloarchaeaviruses is unique for known CCB viromes, the community is still more similar to CCB viral communities than to those of other hypersaline sites, except for other hypersaline microbial mats.

AD are also distinguished from other hypersaline sites by the maintenance of high diversity despite increases in salinity and abundance of haloarchaeaviruses. In fact, AD diversity seems to be higher than in other environments around the world, except for other hypersaline microbial mats and some soda lakes, regardless of the season.

The uniqueness of this viral community is likely related to the great Archaea diversity and virus-host interactions that need further exploration to fully characterize the community dynamics of this exceptional site.

Finally, the similarities between late 2019–2020 surface viromes with depth viromes, which are highly diverse and rich in haloarchaeavirues, supports a hypothesis where hydrological processes such as upwelling of the deep aquifer can function as a 'seed bank' with great microbial diversity.

Funding information
Alejandro Miguel Cisneros-Martínez is a doctoral student from the Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and has received the CONACyT fellowship 814 975. This study was supported by DGAPA/UNAM-PAPIIT (grants numbers IG200319 and IN204822). This manuscript was written with Chile ANID-CEQUA R-20F0009 to VS and LEE as part of larger comparison with ocean waters in the Southern tip of the American continent.

Acknowledgements
We greatly acknowledge Dr Erika Aguirre-Planter, Dr Rosalinda Tapia, Dr Laura Espinosa-Asuar, Jazmín Sánchez-Pérez, Mariette Viladomat-Jasso, Rodrigo Vázquez and Manuel Rosas for their logistical and technical support and assistance during sample collection, and laboratory processing. We also acknowledge Rodrigo García-Herrera for providing access to the LANCIS-Instituto de Ecología's Supercomputer Unit. Finally, we thank SEMARNAT and APFF Cuatro Cienegas for facilitating the sampling and, in particular Rancho Pozas Azules, PRONATURA Noreste for access and permission to sample in the CCB Natural Protected Area.

Author contributions
A.M.C.M., V.S. and L.E.E. were involved in conceptualization, and methodology. A.M.C.M. was responsible for data curation, bioinformatic and statistical analysis, visualization, and writing the original manuscript draft. V.S. and L.E.E. were involved in reviewing and editing, and project supervision. V.S. and L.E.E. acquired 689 funding.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. **Cisneros-Martínez AM, Eguiarte LE, Souza V**. Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat. figshare. Figshare. 2023. https://doi.org/10.6084/m9.figshare.20958184.v1
2. **Souza V, Siefert JL, Escalante AE, Elser JJ, Eguiarte LE**. The Cuatro Ciénegas Basin in Coahuila, Mexico: an astrobiological Precambrian park. *Astrobiology* 2012;12:641–647.
3. **Souza V, Espinosa-Asuar L, Escalante AE, Eguiarte LE, Farmer J, *et al***. An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proc Natl Acad Sci* 2006;103:6565–6570.
4. **Walsh MM**. Microbial mats on the early earth: the archaean rock record. In: **Seckbach J** and **Oren A** (eds). *Microbial Mats: Modern and Ancient Microorganisms in Stratified Ecosystems*. London: Springer; 2010. pp. 43–51.
5. **Alcaraz LD, Olmedo G, Bonilla G, Cerritos R, Hernández G, *et al***. The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proc Natl Acad Sci* 2008;105:5803–5808.
6. **Moreno-Letelier A, Olmedo-Alvarez G, Eguiarte LE, Souza V**. Divergence and phylogeny of *Firmicutes* from the Cuatro Ciénegas Basin, Mexico: a window to an ancient ocean. *Astrobiology* 2012;12:674–684.
7. **Souza V, Moreno-Letelier A, Travisano M, Alcaraz LD, Olmedo G, *et al***. The lost world of Cuatro Ciénegas Basin, a relictual bacterial niche in a desert oasis. *Elife* 2018;7:e38278.

8. **Wolaver BD, Crossey LJ, Karlstrom KE, Banner JL, Cardenas MB,** ***et al***. Identifying origins of and pathways for spring waters in a semiarid basin using He, Sr, and C isotopes: Cuatrocienegas Basin, Mexico. *Geosphere* 2013;9:113–125.
9. **Medina-Chávez NO, Viladomat-Jasso M, Olmedo-Álvarez G, Eguiarte LE, Souza V,** ***et al***. Diversity of archaea domain in Cuatro Cienegas Basin: archaean domes. *bioRxiv* 2019.
10. **Espinosa-Asuar L, Monroy-Guzmán C, Madrigal-Trejo D, Navarro-Miranda M, Sánchez-Pérez J,** ***et al***. Diversity of an uncommon elastic hypersaline microbial mat along a small-scale transect. *PeerJ* 2022;10:e13579.
11. **Madrigal-Trejo D**. Análisis del metametaloma en los tapetes microbianos de Domos del Arqueano, Cuatro Ciénegas, como recapitulación del uso de metales a lo largo de la historia de la Tierra. CdMx (MX): Universidad Nacional Autónoma de México; 2022. http://132.248.9.195/ptd2022/enero/0821431/Index.html [accessed 5 September 2022].
12. **Banciu HL, Sorokin DY**. Adaptation in Haloalkaliphiles and Natronophilic bacteria. In: **Seckbach J, Oren A** and **Stan-Lotter H** (eds). *Polyextremophiles: Life Under Multiple Forms of Stress*. London: Springer; 2013. pp. 123–178.
13. **Halophilic MH**. Acidophilic and Haloacidophilic prokaryotes. In: **Seckbach J, Oren A** and **Stan-Lotter H** (eds). *Polyextremophiles: Life Under Multiple Forms of Stress*. Springer; 2013. pp. 203–213.
14. **Madrigal-Trejo D, Sánchez-Pérez J, Espinosa-Asuar L, Souza V**. Modern microbial mats from the Chihuahuan Desert provide insights into ecological stability throughout Earth's history. *bioRxiv* 2021.
15. **DasSarma S, DasSarma P**. Halophiles. In: *ELS*. Chichester: John Wiley & Sons, 2012.
16. **Ventosa A, Fernández AB, León MJ, Sánchez-Porro C, Rodriguez-Valera F**. The Santa Pola saltern as a model for studying the microbiota of hypersaline environments. *Extremophiles* 2014;18:811–824.
17. **Litchfield CD**. Saline lakes. In: **Reitner J** and **Thiel V** (eds). *Encyclopedia of Geobiology. Encyclopedia of Earth Sciences Series*. Dordrecht: Springer; 2011. pp. 765–768.
18. **Merino N, Aronson HS, Bojanova DP, Feyhl-Buska J, Wong ML,** ***et al***. Living at the extremes: extremophiles and the limits of life in a planetary context. *Front Microbiol* 2019;10:1785.
19. **Boros E, Kolpakova M**. A review of the defining chemical properties of soda lakes and pans: an assessment on a large geographic scale of Eurasian inland saline surface waters. *PLoS One* 2018;13:e0202205.
20. **Kempe S, Kazmierczak J**. Soda lakes. In: **Reitner J** and **Thiel V** (eds). *Encyclopedia of Geobiology. Encyclopedia of Earth Sciences Series*. Dordrecht: Springer; 2011. pp. 824–828.
21. **McGenity TJ, Oren A**. Hypersaline environments. In: **Bell EM** (eds). *Life at Extremes: Environments, Organisms and Strategies of Survival*. Malta: CAB International; 2012. pp. 402–437.
22. **Castelán-Sánchez HG, Elorrieta P, Romoacca P, Liñan-Torres A, Sierra JL,** ***et al***. Intermediate-salinity systems at high altitudes in the Peruvian Andes Unveil a high diversity and abundance of bacteria and viruses. *Genes* 2019;10:891.
23. **Vavourakis CD, Ghai R, Rodriguez-Valera F, Sorokin DY, Tringe SG,** ***et al***. Metagenomic insights into the uncultured diversity and physiology of microbes in four hypersaline Soda Lake Brines. *Front Microbiol* 2016;7:211.
24. **Jones BE, Grant WD, Duckworth AW, Owenson GG**. Microbial diversity of soda lakes. *Extremophiles* 1998;2:191–200.
25. **Roux S, Enault F, Ravet V, Colombet J, Bettarel Y,** ***et al***. Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environ Microbiol* 2016;18:889–903.
26. **Prangishvili D, Bamford DH, Forterre P, Iranzo J, Koonin EV,** ***et al***. The enigmatic archaeal virosphere. *Nat Rev Microbiol* 2017;15:724–739.
27. **Wommack KE, Colwell RR**. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* 2000;64:69–114.
28. **Sullivan MB, Weitz JS, Wilhelm S**. Viral ecology comes of age. *Environ Microbiol Rep* 2017;9:33–35.
29. **Santos F, Meyerdierks A, Peña A, Rosselló-Mora R, Amann R,** ***et al***. Metagenomic approach to the study of halophages: the environmental halophage 1. *Environ Microbiol* 2007;9:1711–1723.
30. **Winter C, Bouvier T, Weinbauer MG, Thingstad TF**. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol Mol Biol Rev* 2010;74:42–57.
31. **Guixa-Boixareu N, Calderón-Paz JI, Heldal M, Bratbak G, Pedrós-Alió C**. Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. *Aquat Microb Ecol* 1996;11:215–227.
32. **Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T,** ***et al***. Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 2008;452:340–343.
33. **Taboada B, Isa P, Gutiérrez-Escolano AL, Del Ángel RM, Ludert JE,** ***et al***. 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* 2018;84:e00465-18.
34. **Alleon J, Bernard S, Olivier N, Thomazo C, Marin-Carbonne J**. Inherited geochemical diversity of 3.4 Ga organic films from the Buck Reef Chert, South Africa. *Commun Earth Environ* 2021;2:6.
35. **Purdy KJ**. Nucleic acid recovery from complex environmental samples. *Methods Enzymol* 2005;397:271–292.
36. **Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB,** ***et al***. Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl Environ Microbiol* 2012;78:6309–6320.
37. **Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M,** ***et al***. Functional metagenomic profiling of nine biomes. *Nature* 2008;452:629–632.
38. **Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F,** ***et al***. Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* 2008;3:e1584.
39. **Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA,** ***et al***. The marine viromes of four oceanic regions. *PLoS Biol* 2006;4:e368.
40. **McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M,** ***et al***. Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS One* 2008;3:e3263.
41. **Roux S, Enault F, Robin A, Ravet V, Personnic S,** ***et al***. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 2012;7:e33641.
42. **Chen S, Zhou Y, Chen Y, Gu J**. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34:i884–i890.
43. **Menzel P, Ng KL, Krogh A**. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 2016;7:11257.
44. **McMurdie PJ, Holmes S**. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8:e61217.
45. **Team RC**. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.
46. **Cameron ES, Schmidt PJ, Tremblay B-M, Emelko MB, Müller KM**. Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Sci Rep* 2021;11:22302.
47. **Wickham H.** ggplot2: elegant graphics for data analysis. New York, USA: Springer; 2016. https://doi.org/10.1007/978-0-387-98141-3
48. **Felsenstein J**. PHYLIP - Phylogeny inference package (version 3.2). *Cladistics* 1989;5:164–166.
49. **Jiménez-Jacinto V, Sanchez-Flores A, Vega-Alvarado L**. Integrative Differential Expression Analysis for Multiple EXperiments (IDEAMEX): a web server tool for integrated RNA-Seq data analysis. *Front Genet* 2019;10:279.

50. **Robinson MD, McCarthy DJ, Smyth GK**. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–140.
51. **Dávila-Ramos S, Castelán-Sánchez HG, Martínez-Ávila L, Sánchez-Carbente MDR, Peralta R, *et al***. A review on viral metagenomics in extreme environments. *Front Microbiol* 2019;10:2403.
52. **Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin EV**. Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res* 2018;244:181–193.
53. **Luk AWS, Williams TJ, Erdmann S, Papke RT, Cavicchioli R**. Viruses of haloarchaea. *Life* 2014;4:681–715.
54. **Pietilä MK, Demina TA, Atanasova NS, Oksanen HM, Bamford DH**. Archaeal viruses and bacteriophages: comparisons and contrasts. *Trends Microbiol* 2014;22:334–344.
55. **González B, Monroe L, Li K, Yan R, Wright E, *et al***. Phage G structure at 6.1 a resolution, condensed DNA, and host identity revision to a *Lysinibacillus*. *J Mol Biol* 2020;432:4139–4153.
56. **Lynch KH, Stothard P, Dennis JJ**. Comparative analysis of two phenotypically-similar but genomically-distinct *Burkholderia cenocepacia*-specific bacteriophages. *BMC Genomics* 2012;13:223.
57. **Nilsson E, Li K, Fridlund J, Šulčius S, Bunse C, *et al***. Genomic and seasonal variations among aquatic phages infecting the Baltic Sea gammaproteobacterium *Rheinheimera* sp. strain BAL341. *Appl Environ Microbiol* 2019;85:e01003-19.
58. **Stroupe ME, Brewer TE, Sousa DR, Jones KM**. The structure of *Sinorhizobium meliloti* phage ΦM12, which has a novel T=19l triangulation number and is the founder of a new group of T4-superfamily phages. *Virology* 2014;450–451:205–212.
59. **Ash KT, Drake KM, Gibbs WS, Ely B**. Genomic diversity of type B3 bacteriophages of *Caulobacter crescentus*. *Curr Microbiol* 2017;74:779–786.
60. **Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, *et al***. Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. *PLoS One* 2012;7:e33802.
61. **Csardi G, Nepusz T**. *The Igraph Software Package for Complex Network Research*. InterJournal; Complex Systems:1695, 2006.
62. **Escalante AE, Eguiarte LE, Espinosa-Asuar L, Forney LJ, Noguez AM, *et al***. Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS Microbiol Ecol* 2008;65:50–60.
63. **García-Ulloa M, Souza V, Esquivel-Hernández DA, Sánchez-Pérez J, Espinosa-Asuar L, *et al***. Recent differentiation of aquatic bacterial communities in a hydrological system in the Cuatro Ciénegas Basin, after a natural perturbation. *Front Microbiol* 2022;13:825167.
64. **Kempe S, Kazmierczak J**. Soda ocean hypothesis. In: **Reitner J** and **Thiel V** (eds). *Encyclopedia of Geobiology. Encyclopedia of Earth Sciences Series*. Dordrecht: Springer; 2011. pp. 829–833.
65. **Soberón J**. Grinnellian and Eltonian niches and geographic distributions of species. *Ecol Lett* 2007;10:1115–1123.
66. **Yachi S, Loreau M**. Biodiversity and ecosystem productivity in a fluctuating environment: the insurance hypothesis. *Proc Natl Acad Sci* 1999;96:1463–1468.
67. **Lanzén A, Simachew A, Gessesse A, Chmolowska D, Jonassen I, *et al***. Surprising prokaryotic and eukaryotic diversity, community structure and biogeography of Ethiopian soda lakes. *PLoS One* 2013;8:e72577.
68. **Gorbushina AA, Krumbein WE**. The poikilotrophic micro-organism and its environment. In: **Seckbach J** (eds). *Enigmatic Microorganisms and Life in Extreme Environments*. Dordrecht: Springer; 1999. pp. 177–185.
69. **Johannesson KH, Cortés A, Kilroy KC**. Reconnaissance isotopic and hydrochemical study of Cuatro Ciénegas groundwater, Coahuila, México. *J South Am Earth Sci* 2004;17:171–180.
70. **Rebollar EA, Avitia M, Eguiarte LE, González-González A, Mora L, *et al***. Water-sediment niche differentiation in ancient marine lineages of Exiguobacterium endemic to the Cuatro Cienegas Basin. *Environ Microbiol* 2012;14:2323–2333.
71. **Delgado-García M, Contreras-Ramos SM, Rodríguez JA, Mateos-Díaz JC, Aguilar CN, *et al***. Isolation of halophilic bacteria associated with saline and alkaline-sodic soils by culture dependent approach. *Heliyon* 2018;4:e00954.
72. **Boyd CE**. pH, carbon dioxide, and alkalinity. In: *Water Quality*. Springer, 2015. pp. 153–178.
73. **Dillon JG**. The role of sulfate reduction in stromatolites and microbial mats: ancient and modern perspectives. In: **Tewari VC** and **Seckbach J** (eds). *STROMATOLITES: Interaction of Microbes with Sediments, Cellular Origin, Life in Extreme Habitats and Astrobiology 18*. Springer; 2011. pp. 571–590.
74. **Lyons WB, Long DT, Hines ME, Gaudette HE, Armstrong PB**. Calcification of cyanobacterial mats in Solar Lake, Sinai. *Geol* 1984;12:623.
75. **Berner RA, Scott MR, Thomlinson C**. Carbonate alkalinity in the pore waters of anoxic marine sediments. *Limnol Oceanogr* 1970;15:544–549.

Artículo 2: Comparative evaluation of bioinformatic tools for virus-host prediction and their application to a highly diverse community in the Cuatro Ciénegas Basin, Mexico

Comparative evaluation of bioinformatic tools for virus-host prediction and their application to a highly diverse community in the Cuatro Ciénegas Basin, Mexico

Alejandro Miguel Cisneros-Martínez[1,2], Ulises E. Rodriguez-Cruz[1,2], Luis D. Alcaraz[3], Arturo Becerra[4], Luis E. Eguiarte[1], Valeria Souza[1,5*]

[1]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Ciudad de México, México

[2]Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México, Ciudad de México, México

[3]Departamento de Biología Celular, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de México, México

[4]Departamento de Biología Evolutiva, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de México, México

[5]Centro de Estudios del Cuaternario de Fuego-Patagonia y Antártica (CEQUA), Punta Arenas, Chile

[*]Corresponding author

E-mail: souza@unam.mx (VS)

Abstract

The sheer diversity of unculturable viruses has prompted the need to describe new viruses through culture-independent techniques. The associated host is one important phenotypic feature that can be inferred from metagenomic viral contigs -- thanks to the development of various bioinformatic tools. Here we compare the performance of recently developed tools for virus-host prediction on a dataset of 1,046 virus-host pairs and then apply the best-performing tools on a metagenomic dataset derived from a highly diverse transiently hypersaline site known as Archaean Domes within the Cuatro Ciénegas Basin, Coahuila, Mexico. We also introduce a virus-host prediction tool called CrisprCustomDB, which uses specific criteria to solve controversial host assignments with custom spacers databases. Host-dependent alignment-based methods showed an average precision of 83% and a sensitivity from 13.7% to 17.7%, whereas host-dependent alignment-free methods achieved an average precision of 75.7% and a sensitivity of 57.5%. RaFAH, a virus-dependent alignment-based tool, had the best performance overall (F1_score = 95.7%). However, when applied to the highly diverse metagenomic dataset, the host-dependent alignment-based (*e.g.*, CrisprCustomDB) and alignment-free (*e.g.*, PHP) methods showed the greatest agreement with each other, even though they are fundamentally different methods. This is because instead of depending on known hosts or viruses-with-known-host databases, they can directly relate metagenomic viral contigs and metagenome-assembled genomes from the same dataset. Such methods also showed the greatest consistency between the source environment and the predicted host taxonomy, habitat, lifestyle, or metabolism, revealing that Archaean Domes viruses likely infect halophilic Archaea as well as a variety of Bacteria which may be halophilic, halotolerant, alkaliphilic, thermophilic, oligotrophic, sulfate-reducing or marine-

related. Consequently, using a combination of methods and qualitative validations relating to the source environment and the predicted host biology will increase the number of correct predictions, mainly when dealing with novel viruses.

Introduction

Until 1990, the International Committee on Taxonomy of Viruses (ICTV) requested detailed information on biological properties to describe and classify new viruses. These biological properties were observed either *in vitro* (*e.g.*, replication cycle, virion structure and antigenic relationships) or in natural host interactions (*e.g.*, pathogenicity, epidemiology, and host range). However, with the development of DNA sequencing techniques, bioinformatics tools, and methods to study molecular evolution, the need to incorporate genomic information --specifically phylogenetic groupings-- to classify novel viruses has arisen [1]. In addition, the development of these tools now allows metagenomic analyses that can detect up to 90% of unknown viruses (non-culturable) in various environments. For example, the TARA ocean expedition assembled viruses from different oceans worldwide and could only assign a family-level classification to 10% of the assembled viral contigs [2]. As a result, the scientific community has proposed incorporating metagenomic data into the ICTV taxonomy by using phenotypic traits and phylogenetic information inferred from assembled sequences and genomes [1].

The need to characterize viruses assembled from metagenomes has prompted various bioinformatic methods for virus-host prediction [3]. A recent study suggests a five-category classification [4] (Fig 1): i) host-dependent alignment-based methods; ii) host-dependent alignment-free methods; iii) virus-dependent alignment-based methods; iv) virus-dependent alignment-free methods; and v) integrative methods. Host-dependent alignment-based methods include methods based on homology signals, such as searching for homology between viral and host proteins, tRNAs, viral genomes and CRISPR spacers, integrated prophages, and protein-protein interactions (PPI). These methods are helpful for detecting recent infections but have the disadvantage that not all viruses share genes with their hosts, which tends to make them precise, but with a low detection rate [3]. CrisprOpenDB is a recently released tool that uses biological criteria to standardize host predictions based on CRISPR spacers with increased sensitivity and precision thanks to its >11 million spacers database derived from >300,000 candidate hosts [5].

Fig 1. Classification of virus-host prediction methods. RaFAH uses host-dependent alignment-based methods to build part of its training database (red discontinuous lines). Integrative methods (iPHoP – blue lines; PHISDetector – green lines; VirHostMatcher-Net – yellow lines) attempt to exploit the virtues of a different number of methods. PPI = protein-protein interactions.

Host-dependent alignment-free methods include those based on sequence composition (*e.g.*, similarity in codon usage, similarity in oligonucleotide composition, and GC content), which rely on the notion that viruses, being genetic parasites, approximate their nucleotide composition to that of the host over time. This genomic mimicry may allow viruses to use the same tRNAs for protein synthesis or to evade the detection and degradation mechanisms of foreign nucleic acids. However, viruses can have similar sequence profiles independently, which can lead to a high false positive rate [3]. VirHostMatcher, which evaluates virus-host genome similarity through d^*_2 distance from 6-mer profiles [6], WIsH, which uses 8-mer profiles and Hidden Markov models (HMMs) [7] and PHP, which uses 4-mer profiles and a Gaussian model [8], are some well-known similar host-dependent alignment-free methods. These alignment-free strategies also include methods based on co-abundance profiles, which rely on the notion that viruses can only be found in the environment in which their host is also found. This profiling method requires the calculation of correlations of normalized abundance profiles of phage and bacteria in different environmental samples. However, they entail a major drawback: predator-prey interactions -- such as those described by the kill-the-winner model [9, 10] -- can generate positive or negative correlations, depending on where the interaction was at the time the sample was taken.

Instead of relying on host databases, virus-dependent methods depend on databases storing viruses with known hosts, to which query viruses are related either through homology signals (alignment-based) or their similarity in oligonucleotide composition (alignment-free). On the one hand, a machine-learning approach named Random Forest Assignment of Hosts (RaFAH) [11] is a virus-dependent alignment-based method that builds a part of its training database from CRISPR spacers, the presence of horizontally transferred genes and common tRNAs to ultimately associate the query virus to a virus with a known host through similarity in protein content. On the other hand, HostPhinder [12] is a virus-dependent alignment-free method that compares 16-mer profiles between query viruses and a database of 2,196 phages with known hosts.

Finally, integrative methods attempt to exploit the virtues of different methods like VirHostMatcher-Net [13], which integrates host-dependent alignment-based methods (CRISPR spacers) and host-dependent alignment-free methods (VirHostMatcher or WIsH) in a network framework, PHISDetector [14], which integrates BLAST [15], CRISPR spacers, prophage, and PPI analyses through a set of machine learning approaches, or iPHoP, which uses machine learning algorithms to compute taxonomy-aware scores for BLAST, CRISPR, VirHostMatcher, WIsH, and PHP, and integrates them with RaFAH results to obtain a final composite score [4].

Optimization of precision and sensitivity estimates within each approach has been achieved by either using more extensive reference databases (*e.g.,* CrisprOpenDB), by leveraging the power of different machine learning algorithms (PHP, RaFAH, VirHostMatcher-Net, PHISDetector, iPHoP), or by integrating different methods (VirHostMatcher-Net, PHISDetector, iPHoP). The publication of these tools is typically accompanied by validation tests with estimates of precision and sensitivity, as well as comparisons with other methods. However, most publications use different databases and sometimes use published values to compare the precision of different methods [6] directly. So far, Roux et al. [4] have compared the largest number of methods showing that host-dependent alignment-based methods can achieve high precision but suffer from low sensitivity. In contrast, host-dependent alignment-free methods have greater sensitivity but struggle to make correct predictions, while virus-dependent alignment-based methods such as RaFAH present both high sensitivity and precision. However, virus-dependent methods may underperform when predicting the host of novel viruses, which also affects, to a lesser extent, host-dependent alignment-based methods but not alignment-free methods [4]. Furthermore, we believe that the fact that host-dependent methods do not necessarily depend on a sizeable pre-compiled reference database (except for CrisprOpenDB) represents an advantage when predicting hosts of novel viruses since they allow to compare metagenomic viral contigs (mVCs) with potential hosts metagenome-assembled genomes (MAGs) from the same environment or even the same dataset.

Here, we present CrisprCustomDB, an algorithm inspired by CrisprOpenDB biological criteria to predict hosts based on CRISPR spacers, which allows the use of custom spacers databases (i.e., spacers predicted from the same metagenomic dataset as mVCs). Furthermore, we evaluated the precision and sensitivity of the different bioinformatic tools for virus-host prediction. Additionally, we applied some of the best performing tools on a metagenomic dataset derived from samples collected at a recently discovered pond known as Archaean Domes in the Cuatro Ciénegas Basin, Coahuila, in the north of Mexico [16, 17]. Archaean Domes is a seasonally fluctuating pond characterized by high pH and salinity, where an extreme diversity of Bacteria and Archaea has been recently described [16, 17]. In addition, its highly diverse viral community does not behave like those from other hypersaline or high pH sites in the face of environmental solid fluctuations [18], for which it is essential to characterize the virus-host relationships and interactions that may drive the microbial and viral diversity in such a unique site.

Materials and Methods

Benchmarking of bioinformatics tools for virus-host prediction

Genomes were selected by downloading three lists (S1_File): i) NCBI complete viral genomes (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi) filtered by host 'bacteria'; ii) Virus-hostDB tabular report (https://www.genome.jp/ftp/db/virushostdb/) and; iii) RefSeq release 217 catalog (https://ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/), filtered by complete genomic molecule, not plasmid.

A link was established between the three tables so that if the viral genome accession had a match in the virus-hostDB, it was checked if the host taxa had a match in the RefSeq catalog (for a given virus with known host, check if the host has a complete genome). 1,029 phage genomes and 133 bacterial genomes were downloaded from NCBI (S2_File), which together made up 1,046 virus-host pairs with complete genomes. The performance of the virus-host prediction tools was evaluated at the genus level.

A custom script was written (https://github.com/AleCisMar/CrisprCustomDB/blob/main/benchmarking/compare_real-estimated.pl) to compare the estimated virus (accession)-host (genus) pairs with the actual virus (accession)-host (genus) pairs in order to obtain the true positives (TP), false positives (FP) and false negatives (FN) for every prediction tool. TP is the successful rejection of a null hypothesis denying any relationship between the real pairs. FP would typically be the incorrect rejection of a second null hypothesis about the relationship between falsely predicted or observed pairs (type I error). However, given that prediction tools are tested against a reference list of confirmed or expected virus-host pairs in this particular evaluation, FP is also incurring a type II error, which is the failure to reject that there is no relationship between the real pairs. Thus, FN includes FP and viruses with unassigned hosts (NA) (type II error).

Precision, sensitivity, and F1_score were calculated as follows:

Precision (Positive Predictive Value):

$$PPV = \frac{TP}{TP + FP}$$

Sensitivity (True Positive Rate):

$$TPR = \frac{TP}{TP + FN}$$

F1_score:

$$F1_{score} = 2\left(\frac{PPV * TPR}{PPV + TPR}\right)$$

To run each program, we followed the instructions provided by the developers choosing the parameters for which they are reported to perform the best [5, 6, 7, 8, 11, 12, 13]. In addition, a Perl script inspired by CrisprOpenDB was developed (CrisprCustomDB available at https://github.com/AleCisMar/CrisprCustomDB). CrisprCustomDB uses the host assignment criteria of Dion et al. [5]. These are: (i) host if spacers have a maximum of 2 mismatches with the viral genome; (ii) if multiple candidates meet criteria 1, the one with the most spacers aligning with different regions of the viral genome is selected; (iii) if there are multiple candidates meeting criterion 2, the one with the spacer closest to the 5' end in the array is selected; and (iv) if there are multiple candidates meeting criteria three the lowest shared taxonomic rank is assigned.

The CRISPR spacers were obtained with the CRISPRDetect tool [19] to test their performance on the custom database (133 bacterial genomes) with the parameters recommended by Dion et al. [5]. Since only 1,349 spacers (S2_File) were found in 40 of the 133 bacterial genomes (30%), two sensitivity calculations were made, one considering the 1,046 virus-host pairs and the other taking into account only the maximum number of virus-host pairs (261) that can be obtained given the number of possible hosts with spacers (40). CrisprCustomDB, VirHostMatcher, WIsH, and PHP were run on the custom database (1,029 phage genomes and 133 bacterial genomes).

Since HostPhinder, CrisprOpenDB, VirHostMatcher-Net, and RaFAH rely on large reference databases, only the 1,029 phage genomes were used as input. For PHP, an additional estimation was made using the reference database with 60,105 potential hosts provided by the authors. For VirHostMatcher, two estimations were made, both with a score

≤ 0.25. The first selects the most frequent host within the top 30 with the most similar profiles, and the second selects the most frequent host within the top 5 with the most similar profiles. For VirHostMatcher-Net, two estimations were also made. One without score restriction and the other limited to predictions with a score > 0.95.

Virus-host prediction on assembled metagenomic reads

Sample collection and sequencing

Sampling was carried out at the Archaean Domes of the Rancho Pozas Azules (26°49'41.9" N, 102°01'23.6" W), belonging to Pronatura Noreste, in the Cuatro Ciénegas Basin, Coahuila, in the North of Mexico, under SEMARNAT scientific permit number SGPA/DGVS/03121/15. Twelve samples were taken between 2016 and 2020. For microbial mats, seven surface samples (M1 – M6 and D0) were collected using a sterile scalpel dissection (8 cm^2 / 40 cm^3) and transferred to 50 mL conical tubes. 30 cm plastic tubes were used as sediment samplers to collect two additional microbial mat samples at 30 and 50 cm depth (D30 and D50). Three samples were collected at the shallow ellipsoid orange pools or orange circles (OC) [16, 17, 18]: one superficial water sample (C0) on a 50 mL conical tube and two more at depths of 30 and 50 cm (C30 and C50). All samples were stored in liquid nitrogen until processing.

DNA was extracted according to [20] at the Laboratorio de Evolución Molecular y Experimental of the Instituto de Ecología, Universidad Nacional Autónoma de México, in Mexico City. Briefly, the extractions followed a column-based protocol with a Fast DNA Spin Kit for Soil (MP Biomedical) [21]. Total DNA was sent to CINVESTAV-LANGEBIO, Irapuato, México, for shotgun sequencing with Illumina Mi-Seq paired-end 2x300 technology.

All sequence reads are available on the National Centre for Biotechnology Information (NCBI) Sequence Reads Archive (SRA) under the BioProject accession: PRJNA847603.

Read processing and assembly of mVCs

The read quality was assessed with FastQC v0.11.9 [22]. Adapter removal and quality filtering were performed with Trimmomatic v0.39 [23] using a sliding window of 4 base pairs excluding reads with an average quality of less than 30 and less than 20 nucleotides. Clean reads were assembled with SPAdes 3.15.2 [24] using the --metaviral option. The viralVerify and viralComplete scripts (included in the SPAdes package) were used to verify that the assembled contigs correspond to viral genomes and to assess genome completeness, respectively. The circularity of the viral contigs was checked. When necessary, the position of sequences was adjusted prior to gene prediction and annotation with the help of custom scripts (available at https://github.com/AleCisMar/GenomicTools) that make use of BLAST [15], EMBOSS [25], Prodigal [26], and HMMER [27].

Read processing, assembly, and taxonomic assignment of MAGs

The quality of the raw data was assessed with FastQC (v0.11.8) [22] and filtered with Trimmomatic (v0.39) [23]. The reads were then assembled using MetaSPAdes (v3.15.3) [28], and the contigs obtained in the assembly were used to perform read binning or clustering, which was performed with MaxBin2 (v2.2.7) [29] and MetaBat2 (v2.12.1) [30].

Binning refiner (v1.4.2) software [31] was used to reduce the percentage of contamination in the bins. The integrity of the MAGs was assessed using CheckM (v1.1.3) [32] with the default settings.

For taxonomic assignment and placement of MAGs on the phylogenetic tree of life, we used the program GTDB-tk (v1.6.0) [33], which identifies 122 and 120 marker genes of archaea and bacteria, respectively, using HMMER [27]. Briefly, genomes are assigned to the domain with the most identified marker genes. Selected domain-specific markers are aligned with HMMER, concatenated into a single multiple sequence alignment, and trimmed with the ~5000-column Bacteria or Archaea mask used by GTDB [33].

Implementation of virus-host prediction tools on metagenomic data

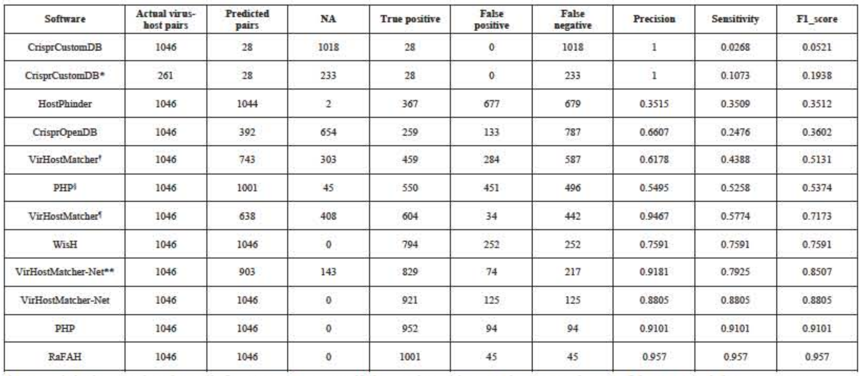
After taxonomic classification of the MAGs, prediction of the spacer sequences of the CRISPR arrays found in the MAGs was performed using the CRISPRCasTyper program (v 1.3.0) [34] using the following parameters cctyper -t 4 --prodigal single –circular. 2,660 spacers (S3_File) were found. The mVCs were run against the spacer database with Blastn [15], allowing for a maximum of 2 mismatches. Spacers were predicted with the CRISPRDetect tool [19] to implement CrisprCustomDB pipeline using array_quality_score_cutoff of 3, as recommended for FASTA files. This spacers prediction for CrisprCustomDB analysis resulted in 1,062 spacers (S3_File). The mVCs were also run with CrisprOpenDB, which uses an 11,674,395 spacers database [5]. Virus-host predictions were also made with RaFAH [11] and PHP [8]. For PHP, k-mer frequencies were calculated for all MAGs (S3_File).

Results

Benchmarking of bioinformatics tools for virus-host prediction

The best three performing tools for complete bacteria and phage genomes datasets (F1_score) were RaFAH, PHP, and VirHostMatcher-Net. They were followed by WIsH, VirHostMatcher, CrisprOpenDB, HostPhinder, and CrisprCustomDB at the bottom (Table 1).

Table 1. Precision, sensitivity, and F1_score estimates of the different virus-host prediction tools.



Software	Actual virus-host pairs	Predicted pairs	NA	True positive	False positive	False negative	Precision	Sensitivity	F1_score
CrisprCustomDB	1046	28	1018	28	0	1018	1	0.0268	0.0521
CrisprCustomDB*	261	28	233	28	0	233	1	0.1073	0.1938
HostPhinder	1046	1044	2	367	677	679	0.3515	0.3509	0.3512
CrisprOpenDB	1046	392	654	259	133	787	0.6607	0.2476	0.3602
VirHostMatcher†	1046	743	303	459	284	587	0.6178	0.4388	0.5131
PHP§	1046	1001	45	550	451	496	0.5495	0.5258	0.5374
VirHostMatcher¶	1046	638	408	604	34	442	0.9467	0.5774	0.7173
WisH	1046	1046	0	794	252	252	0.7591	0.7591	0.7591
VirHostMatcher-Net**	1046	903	143	829	74	217	0.9181	0.7925	0.8507
VirHostMatcher-Net	1046	1046	0	921	125	125	0.8805	0.8805	0.8805
PHP	1046	1046	0	952	94	94	0.9101	0.9101	0.9101
RaFAH	1046	1046	0	1001	45	45	0.957	0.957	0.957

*Sensitivity calculated from 261 possible pairs given the number of hosts with spacers.

†Prediction using score ≤ 0.25 and selecting the most frequent host among top 30.

§Using PHP reference database with 60,105 prokaryotic genomes.

¶Prediction using score ≤ 0.25 and selecting the most frequent host among top 5.

**Prediction using score > 0.95.

CrisprOpenDB made 392 predictions, of which 259 were correctly estimated. These results translate into a sensitivity of 24.76%, a precision of 66.07%, and an F1_score of 36.02% (Fig 2). On the other hand, CrisprCustomDB only predicted 28 pairs, all of which were correct (precision = 100%). Considering that we are benchmarking on 1,046 virus-host pairs, CrisprCustomDB reached a sensitivity of only 2.68% and an F1_score of 5.21%. It is important to note that spacers were identified in only 30% of the bacterial genomes in the custom database. Consequently, the maximum number of predicted pairs was 261. This consideration increases the sensitivity to 10.73% and the F1_score to 19.38% (Fig 2).

Fig 2. Precision, sensitivity, and F1_score estimates of the different virus-host prediction tools. For CrisprCustomDB*, sensitivity was estimated considering 261 possible pairs. VirHostMatcher† was tested with a score ≤ 0.25 and selected the most frequent host within the top 30. VirHostMatcher¶ was tested with the same parameters but selecting the most frequent host within the top 5. PHP§ was tested against a reference database of 60,105 potential hosts. For VirHostMatcher-Net**, only predictions with a score > 0.95 were kept.

Alignment-free methods evaluated here make predictions by comparing the oligonucleotide profile of a virus to either the oligonucleotide profile of viruses with a known host (HostPhinder) or the oligonucleotide profile of bacteria (VirHostMatcher, WIsH, PHP). Although HostPhinder predicted 1,044 pairs, most predictions were incorrect (677). Hence, it had the lowest performance of the alignment-free methods (Fig 2), with a sensitivity of 35.09%, a precision of 35.15%, and an F1_score of 35.12%.

VirHostMatcher was executed with two different criteria: i) selecting the most frequent host among the top thirty and; ii) selecting the most frequent host among the top five. When using the first criterion, VirHostMatcher generated more predictions (743 compared to 638) and produced more false positives (284 compared to 34). As a result, it achieved lower sensitivity (43.88% compared to 57.74%), precision (61.78% compared to 94.67%), and F1_score (51.31% compared to 71.73%).

Among these methods, WIsH and PHP emerged as the top predictors, achieving the maximum number of pairs (1,046). WIsH demonstrated a sensitivity, precision, and F1_score of 75.91%, whereas PHP appeared as the best-performing alignment-free method (Fig 2) with a sensitivity, precision, and F1_score of 91.01%. PHP was also tested against a reference database with 60,105 potential hosts provided by the authors. However, this test resulted in fewer predictions (1,001) and lower sensitivity (52.58%), precision (54.95%), and F1_score (53.74%).

VirHostMatcher-Net was executed using two approaches: first, by setting a prediction threshold with a score > 0.95 and, second, without any score restrictions. Restricting the final host assignment to predictions with higher scores resulted in higher accuracy (91.81% vs. 88.05%) at the expense of lower sensitivity (79.25% vs. 88.05%) and, as a consequence, a lower F1 score (85.07% vs. 88.05%). Meanwhile, RaFAH achieved an accuracy, sensitivity, and F1 score of 95.70%, making it the algorithm with the best overall performance (Fig 2).

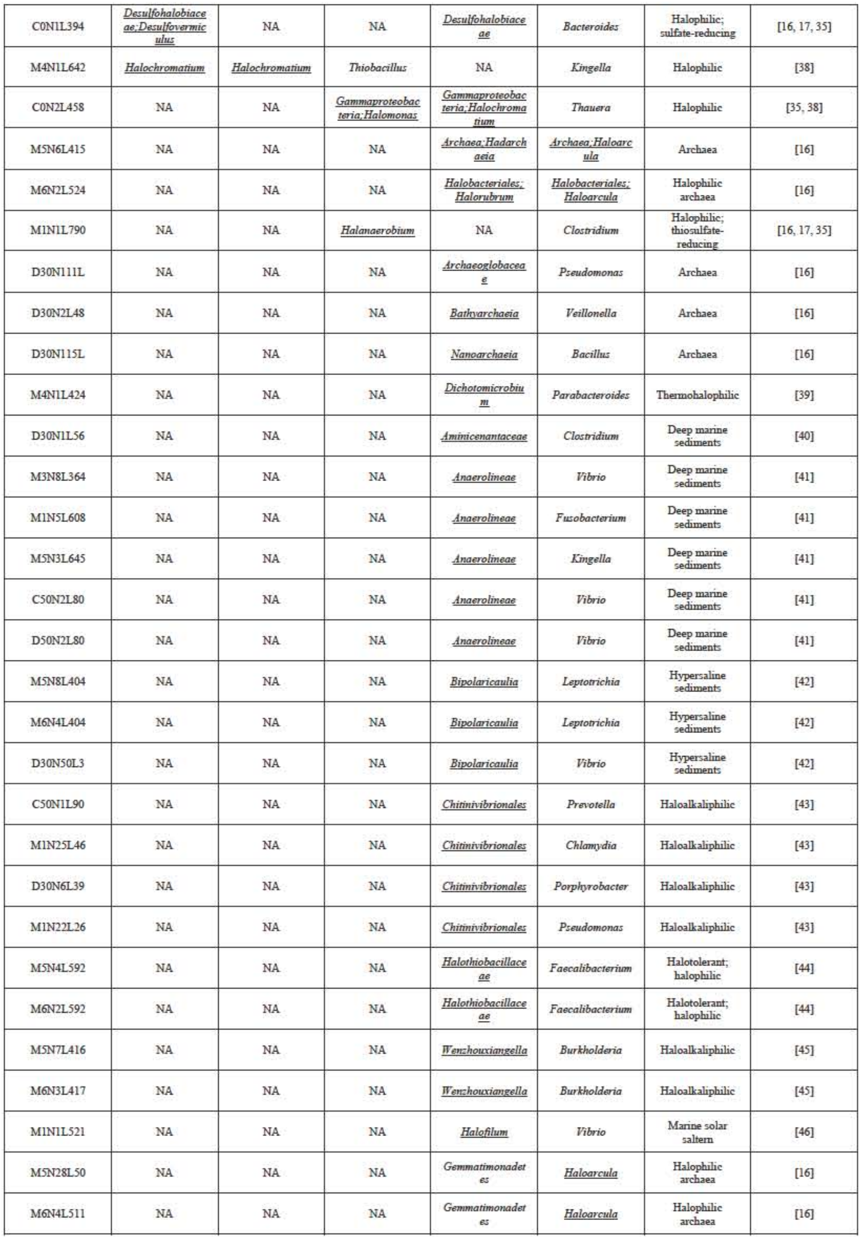
Virus-host predictions on assembled metagenomic reads from Archaean Domes, Cuatro Ciénegas Basin, Mexico

To predict the host of mVCs from Archaean Domes at Cuatro Ciénegas Basin, based on CRISPR spacers predicted on MAGs from the same dataset, we employed two related approaches. The first approach involved conducting a Blastn search using 2,660 spacers, with a maximum of 2 mismatches as the only criterion. The second approach involved CrisprCustomDB, using 1,062 spacers to solve problematic host assignments. Additionally, we performed predictions using CrisprOpenDB, PHP, and RaFAH, as these tools demonstrated superior performance in their respective categories. Since HostPhinder showed lower performance than all other alignment-free methods, and PHP and RaFAH outperformed VirHostMatcher-Net, we did not make predictions with these tools. While PHP was executed on the Archaean Domes MAGs, CrisprOpenDB and RaFAH only required the mVCs, as they relied on their reference databases.

The ordinary CRISPR approach resulted in eight predictions (Table 2). Half of the mVCs (C50N1L42, C0N5L506, M1N5L607, and C9N1L394) were assigned to hosts in the phylum *Desulfobacterota*. The other half were associated with bacteria of the class *Gammaproteobacteria.* Two mVCs (M5N2L438 and M6N1L439) were predicted to infect bacteria of the genus *Halorhodospira*. One contig, the M4NL642, was assigned to the genus *Halochromatium,* and the contig C30N1L64, was assigned to two possible hosts: *Thiohalorhabdus* or *Thiohalospira*.

Table 2. 46 host predictions on mVCs from Archaean Domes Pond, Cuatro Ciénegas, Mexico, designated as reliable according to different criteria.

Contig	CRISPR	CrisprCustomDB	CrisprOpenDB	PHP	RaFAH	Supporting evidence	Reference
C50N1L42	*Desulfovibrionales;Desulfovermiculus*	NA	NA	*Desulfovibrionales;Desulfohalobiaceae*	*Desulfovibrionales;Desulfovibrio*	Halophilic; sulfate-reducing	[16, 17, 35]
M5N2L438	*Halorhodospira*	*Halorhodospira*	NA	*Halorhodospira*	*Pseudomonas*	Halophilic	[36]
M6N1L439	*Halorhodospira*	*Halorhodospira*	NA	*Halorhodospira*	*Pseudomonas*	Halophilic	[36]
C30N1L64	*Thiohalorhabdus/ Thiohalospira*	*Thiohalorhabdus**	NA	*Thiohalorhabdus*	*Vibrio*	Halophilic; sulfur-oxidizing	[35, 37]
C0N5L506	*Desulfobacterales*	NA	NA	*Desulfobacterales*	*Clostridium*	Sulfate-reducing	[16]
M1N5L607	*Desulfobacterales*	*Desulfobacterales*	NA	NA	*Pseudoalteromonas*	Sulfate-reducing	[16]



C0N1L394	*Desulfohalobiaceae;Desulfovermiculus*	NA	NA	*Desulfohalobiaceae*	*Bacteroides*	Halophilic; sulfate-reducing	[16, 17, 35]
M4N1L642	*Halochromatium*	*Halochromatium*	*Thiobacillus*	NA	*Kingella*	Halophilic	[38]
C0N2L458	NA	NA	*Gammaproteobacteria;Halomonas*	*Gammaproteobacteria;Halochromatium*	*Thauera*	Halophilic	[35, 38]
M5N6L415	NA	NA	NA	*Archaea;Hadarchaeia*	*Archaea;Haloarcula*	Archaea	[16]
M6N2L524	NA	NA	NA	*Halobacteriales; Halorubrum*	*Halobacteriales; Haloarcula*	Halophilic archaea	[16]
M1N1L790	NA	NA	*Halanaerobium*	NA	*Clostridium*	Halophilic; thiosulfate-reducing	[16, 17, 35]
D30N111L	NA	NA	NA	*Archaeoglobaceae*	*Pseudomonas*	Archaea	[16]
D30N2L48	NA	NA	NA	*Bathyarchaeia*	*Veillonella*	Archaea	[16]
D30N115L	NA	NA	NA	*Nanoarchaeia*	*Bacillus*	Archaea	[16]
M4N1L424	NA	NA	NA	*Dichotomicrobium*	*Parabacteroides*	Thermohalophilic	[39]
D30N1L56	NA	NA	NA	*Aminicenantaceae*	*Clostridium*	Deep marine sediments	[40]
M3N8L364	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
M1N5L608	NA	NA	NA	*Anaerolineae*	*Fusobacterium*	Deep marine sediments	[41]
M5N3L645	NA	NA	NA	*Anaerolineae*	*Kingella*	Deep marine sediments	[41]
C50N2L80	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
D50N2L80	NA	NA	NA	*Anaerolineae*	*Vibrio*	Deep marine sediments	[41]
M5N8L404	NA	NA	NA	*Bipolaricaulia*	*Leptotrichia*	Hypersaline sediments	[42]
M6N4L404	NA	NA	NA	*Bipolaricaulia*	*Leptotrichia*	Hypersaline sediments	[42]
D30N50L3	NA	NA	NA	*Bipolaricaulia*	*Vibrio*	Hypersaline sediments	[42]
C50N1L90	NA	NA	NA	*Chitinivibrionales*	*Prevotella*	Haloalkaliphilic	[43]
M1N25L46	NA	NA	NA	*Chitinivibrionales*	*Chlamydia*	Haloalkaliphilic	[43]
D30N6L39	NA	NA	NA	*Chitinivibrionales*	*Porphyrobacter*	Haloalkaliphilic	[43]
M1N22L26	NA	NA	NA	*Chitinivibrionales*	*Pseudomonas*	Haloalkaliphilic	[43]
M5N4L592	NA	NA	NA	*Halothiobacillaceae*	*Faecalibacterium*	Halotolerant; halophilic	[44]
M6N2L592	NA	NA	NA	*Halothiobacillaceae*	*Faecalibacterium*	Halotolerant; halophilic	[44]
M5N7L416	NA	NA	NA	*Wenzhouxiangella*	*Burkholderia*	Haloalkaliphilic	[45]
M6N3L417	NA	NA	NA	*Wenzhouxiangella*	*Burkholderia*	Haloalkaliphilic	[45]
M1N1L521	NA	NA	NA	*Halofilum*	*Vibrio*	Marine solar saltern	[46]
M5N28L50	NA	NA	NA	*Gemmatimonadetes*	*Haloarcula*	Halophilic archaea	[16]
M6N4L511	NA	NA	NA	*Gemmatimonadetes*	*Haloarcula*	Halophilic archaea	[16]

C0N2L195	NA	NA	NA	*Halanaerobiales*	*Alistipes*	Halophilic; thiosulfate-reducing	[16, 17, 35]
M4N3L527	NA	NA	NA	*Phycisphaerales*	*Pseudomonas*	Marine	[47]
M1N3L461	NA	NA	NA	*Rhodothermales*	*Alistipes*	Thermohalophilic ; haloalkaliphilic	[48, 49]
D30N26L5	NA	NA	NA	*Petrotogales*	*Fusobacterium*	Thermophilic	[50]
C0N1L567	NA	NA	NA	*Halanaerobiales*	*Clostridium*	Halophilic	[51]
C30N1L45	NA	NA	NA	NA	*Salinispora*	Marine sediments	[52]
M1N6L535	NA	NA	NA	NA	*Desulfotomaculum*	Thermophilic; sulfate-reducing	[53]
M1N8L483	NA	NA	NA	NA	*Thermus*	Thermophilic	[54]
M5N19L71	NA	NA	NA	NA	*Caulobacter*	Oligotrophic	[55]
M6N6L714	NA	NA	NA	NA	*Caulobacter*	Oligotrophic	[55]

Reliable predictions, either through consistency between methods or consistency between the source environment and the predicted host biology (taxonomy, habitat, lifestyle, or metabolism), are underlined. Where applicable, the lowest common taxonomic rank and the lowest taxonomic rank achieved by each tool are separated by ";". The full list of predictions can be found in the S4_File.

*Assigned using criterion 3: Multiple hosts matching the same number of regions. Host with spacer closest to the 5' end.

CrisprCustomDB made five predictions, all consistent with those made by the ordinary CRISPR approach. These included one of the *Desulfobacterota* and three *Proteobacteria*. The fifth prediction assigned contig C30N1L64 to *Thiohalorhabdus* for being the host with the spacer closest to the 5' end, solving the problem of two possible hosts (Table 2). CrisprOpenDB made five predictions (S4_File). The only contig for which all three CRISPR-based methods made a prediction is contig M4N1L642. However, ordinary CRISPR and CrisprCustomDB predicted it to infect the proteobacteria *Halochromatium*, while CrisprOpenDB predicted *Thiobacillus* as the phage host. CrisprOpenDB predicted contig C0N2L458 to infect *Gammaproteobacteria* of the genus *Halomonas* (Table 2).

PHP made 54 predictions. Three of the contigs assigned to *Desulfobacterota* by the ordinary CRISPR approach (C50N1L42, C0N5L506, and C9N1L394) were equally assigned by PHP. Also, three contigs assigned to *Proteobacteria* in both the ordinary CRISPR and CrisprCustomDB approaches were independently assigned the same by PHP. This concordance includes contigs M5N2L438 and M6N1L439, assigned to *Halorhodospira*, and contig C30N1L64, which was also assigned to *Thiohalorhabdus*. Additionally, PHP agreed with CrisprOpenDB on the host assignment for contig C0N2L458 at the class level but suggestd it infects bacteria of the genus *Halochromatium* instead of *Halomonas* (Table 2).

RaFAH produced 87 predictions, of which only three were supported by the other methods. These included the host assignment for contig C50N1L42 (*Desulfobacterota*), which is consistent with the ordinary CRISPR approach and PHP, and the host assignment for contig M5N6L415, which is consistent with PHP in predicting it to infect Archaea. The most similar assignment was observed for contig M6N2L524, predicted to infect *Euryarchaeota* of the genus *Haloarcula* by RaFAH and *Euryarchaeota* of the genus *Halorubrum* by PHP (Table 2). Finally, RaFAH was the only method that correctly predicted the host of *Escherichia* virus ΦX174, which was used as a positive control for DNA sequencing (S4_File).

Discussion

The increasing number of virus-host prediction tools prompted us to perform a comparative evaluation of the most popular and recently released tools (Fig 1). Unfortunately, due to computational limitations related to the size of the databases, we could not evaluate either PHISDetector [14] or iPHoP [4]. Since there is a discrepancy in the performance of PHISDetector compared to VirHostMatcher-Net [4, 14], we can only conclude which tool performs the best once we compare them under the same methodological framework. As for iPHoP, this is probably the best-performing integrative tool [4], as it integrates RaFAH into its host prediction algorithm, which has shown better performance than VirHostMatcher-Net both here (Fig 2) and in its original publication [11].

According to the literature, it is understood that following iPHoP, PHISDetector (compared with VirHostMatcher-Net, PHP, WIsH and VirHostMatcher) [14] and RaFAH (reported with higher F1 score than the combination of CRISPR, BLAST and tRNAs, followed by VirHostMatcher-Net, WIsH, HostPhinder, and CRISPR, BLAST and tRNAs individually) [11] are the most precise tools. They are likely to be followed by VirHostMatcher-Net (more precise than similarity networks, CRISPR, BLAST, WIsH, and VirHostMatcher) [13], PHP (reported less precise than CRISPR and BLAST, which, however, have very low sensitivity, but are more precise than WIsH and VirHostMatcher) [8] and WIsH (reported to be more precise than VirHostMatcher, especially for incomplete or short viral genomes) [7]. Lastly, CrisprOpenDB (reported to have similar precision to WIsH) [5], HostPhinder (reported to be more precise than BLAST) [12], and VirHostMatcher (compared to values published by Edwards et al. [3] appears to have similar precision to homology methods (BLAST, prophage, and CRISPR) and higher than early implementations of the k-mer method, abundance profiling and GC content) [6] appear to be the least precise tools.

To test the above interpretations about the performance of virus-host prediction tools, we downloaded 1,029 and 133 complete phages and bacterial genomes, respectively.

(S1_File and S2_File), making up 1,046 virus-host pairs. We did not use Archaea viruses and their respective hosts, because we could only retrieve eight pairs following the method described in the Materials and Methods section. In addition, some of the virus-host prediction tools evaluated here are explicitly trained on Bacteria and their corresponding phages (*e.g.* [5]) and, therefore, cannot be used to evaluate thier performance on viruses of Archaea. The performance of the virus-host prediction tools was evaluated at the genus level because performance comparisons are often consistent across taxonomic ranks [6, 7, 8, 11, 13, 14], and because it may be more biologically informative than higher taxonomic rank predictions.

As noted elsewhere [3, 4, 8], CRISPR-based methods demonstrated high precision at the expense of sensitivity. In our study, CrisprCustomDB achieved 100% precision but only made 28 out of 1,046 predictions, resulting in extremely low sensitivity. This limitation may be attributed to only 30% of the bacterial genomes having spacers, automatically excluding the remaining 70% from host predictions. Even when considering the maximum number of virus-host pairs that can be predicted given the number of hosts with spacers, sensitivity remained the lowest of all compared methods. Alternatively, the ability of CRISPR-based methods to detect virus-host pairs may be hampered by the host sequence selection process (see Materials and Methods), which excludes plasmid sequences and, therefore, any CRISPR-Cas system possibly encoded therein [56].

As expected, the limited sensitivity of the CRISPR-based methods was overcome by CrisprOpenDB due to the use of a > 11 million spacers database, which made 364 more predictions while retaining a precision higher than some sequence compositions methods, making it a reliable alternative when only viral contigs are available. Although CrisprOpenDB has increased the sensitivity of CRISPR-based methods, they still need to catch up to newer sequence composition methods, which exhibit high sensitivity and improved precision. Such was the case of VirHostMatcher, WIsH, and PHP, which achieved sensitivity and precision > 50%. In contrast, HostPhinder had a higher sensitivity than CRISPR-based methods but the lowest precision among all compared methods. This result suggests that relying solely on transferring the host of the most similar virus may be a greedy and unreliable approach, especially when dealing with a highly diverse viral community with many unknown viruses.

VirHostMatcher did not perform better when assigning the most frequent taxon among a more significant number of possible hosts (up to 30) with a score ≤ 0.25, contrary to what has been reported [6]. Instead, using this consensus criterion among the top 30 scoring hosts yielded a precision even lower than that of CrisprOpenDB and WIsH, which is known to perform better with incomplete contigs [7], while assigning host among the top 5 reached the second highest precision overall. Such discrepancies may depend on the distribution of taxa within the studied dataset. For instance, while increasing the n possible hosts criterion, one can expect a higher probability of finding multiple high-scoring instances of a particular host only by chance on a highly diverse dataset.

PHP allows predictions to be made with custom databases and provides a database of 60,105 bacterial genomes within the program's repository. Using this reference database, PHP obtains the second-lowest precision overall, while the custom database (133 bacterial genomes) elevated PHP as the most accurate and sensitive sequence composition method. This result implies that using an extensive reference database does not necessarily enhance the performance of virus-host prediction tools, unless the actual hosts are present. Thus, PHP may be a suitable tool, especially when working with MAGs and mVCs from the same metagenome. Also, although not directly tested, host-dependent alignment-free methods such as PHP were noticeably more effortless to set up and faster to execute than integrative methods and virus and host-dependent alignment-based methods.

RaFAH achieved the highest precision, sensitivity, and F1_score on the test data collection. However, only a couple of its predictions on the metagenomic dataset were consistent with those of CRISPR-based methods, PHP, or the environment from which the metagenomes were generated. The metagenomic data analyzed here came from samples taken within the Cuatro Ciénegas Basin which, despite being a desert oasis with oligotrophic waters, is known for sheltering diverse groups of microorganisms, many of which are endemic and related to marine microorganisms [57, 58]. Such diversity is believed to have evolved as a result of the long-standing environmental stability of a deep aquifer that recreates an ancient ocean conditions, and which nourishes the aquatic systems of Cuatro Ciénegas Basin through the movement of groundwater produced by the magmatic pouch deep in the Sierra San Marcos y Pinos [59]. Specifically, the environment from which samples were extracted is a shallow pond characterized by high pH and salinity known as Archaen Domes [16, 17, 18]. It has been shown that Archaean Domes harbors a great diversity of bacteria on a short spatial scale [17] and is one of the most diverse archaeal communities in the world [16]. Such diversity includes sulfate-reducing *Proteobacteria* and extreme halophilic *Euryarchaeota* [35]. In addition, a highly diverse viral community has recently been described where haloarchaeaviruses constitute an essential part [18]. Therefore, predictions pointing to halophilic Archaea, as well as halophilic, halotolerant, alkaliphilic, thermophilic, oligotrophic, sulfate-reducing, sulfur-oxidizing or marine Bacteria, were considered consistent with the environment in question (Table 2).

Although CrisprCustomDB was able to discriminate between possible hosts for contig C30N1L64 (further supported by PHP), the fact that the ordinary CRISPR approach made more predictions on the metagenomic dataset than CrisprCustomDB is likely reflecting the benefit of using a more extensive spacer database (see Materials and Methods) as previously discussed regarding the performance of CrisprOpenDB. However, the lack of consistency of CrisprOpenDB and RaFAH with the other methods suggests that relying on a >11 million spacers database [5] or on a Random Forest classifier based on the protein content of viruses with known host [11], respectively, may be beneficial only when the hosts or the assembled viruses are already known, or are closely related to hosts or viruses represented in the corresponding databases. Therefore, for highly diverse datasets likely to have a high proportion of novel viruses as the one tested here [18], it may be more appropriate to use host-based tools, either alignment-based or alignment-free, such as CrisprCustomDB or PHP, with *ad hoc* databases built with archaea and bacteria MAGs from the same dataset whenever possible.

Predictions on the metagenomic dataset show that fundamentally different methods such as CrisprCustomDB and PHP, can complement and support each other. Incorporating these tools along with RaFAH, the best-performing tool on the test dataset, in an integrative software such as iPHoP [4], allows tackling the host prediction problem from different angles, increasing the chance of making the correct predictions. Also, judging the predictions based on the consistency between the predicted host biology (*i.e.*, taxonomy, habitat, lifestyle, or metabolism) and the source environment of the query virus (Table 2) may provide additional validation, mainly when predicting hosts of novel viruses. However, some caution still needs to be exercised with this validation approach. For example, for predictions with less consistency between methods and at higher taxonomic ranks, there is an increased risk that the consistency between the source environment of mVCs and the biology of the predicted hosts will be rather ambiguous or even false.

Host prediction is one of the most critical features for characterizing mVCs, probably along with phylogenetic relationships. We wanted to know who the host is to learn more about the biology of the newly assembled virus, such as where it gets the resources to complete its replication cycle, what organisms it interacts with, and with whom it might co-evolve. However, although the host predictions presented here take us a step forward in characterizing Archaean Domes viruses, we still need to know the phylogenetic context, the evolutionary processes, and the functional adaptations that will allow us to better understand the origin of diversity at this particular site.

Conclusions

The results presented here indicated that RaFAH, a virus-dependent alignment-based method, and PHP, a host-dependent alignment-free method, are the best-performing tools for virus-host prediction. Other methods showed different performances depending on the host selection criteria, scoring thresholds, and the reference database. It seems that CRISPR-based methods seem to benefit from using a more extensive spacers database when predicting hosts of already-known viruses. However, using a more extensive candidate host database did not enhance the performance of host-dependent alignment-free methods such as PHP.

The complementarity and support shown by CrisprCustomDB and PHP when executed on mVCs and MAGs from the same dataset, suggest that using such a combination of tools along with RaFAH may produce more reliable host assignments on highly diverse metagenomic datasets provided that predictions are consistent across multiple methods and the predicted host taxonomy, habitat, lifestyle, or metabolism is consistent with the source environment.

Finally, host predictions on mVCs from Archaean Domes showed that viruses inhabiting such environment infect halophilic Archaea as well as a variety of Bacteria which may be halophilic, halotolerant, alkaliphilic, thermophilic, oligotrophic, sulfate-reducing or marine-related. These predictions are consistent with the particular environment and the known geological and biological evolution of the Cuatro Ciénegas Basin and its microorganisms.

Acknowledgments

We greatly acknowledge Dr. Erika Aguirre-Planter, Dr. Rosalinda Tapia, Dr. Laura Espinosa-Asuar, Jazmín Sánchez-Pérez, Mariette Viladomat-Jasso, Rodrigo Vázquez, and Manuel Rosas from the Instituto de Ecología, UNAM, for their logistical and technical support and assistance during sample collection, and laboratory processing. We also acknowledge Dr. José Alberto Campillo-Balderas for reviewing the manuscript, and Rodrigo García-Herrera for providing access to the LANCIS-Instituto de Ecología's Supercomputer Unit at UNAM. Finally, we thank SEMARNAT and APFF Cuatro Ciénegas for facilitating the sampling and, in particular, Rancho Pozas Azules, PRONATURA Noreste for access and permission to sample in the Cuatro Ciénegas Basin Natural Protected Area.

References

1. Simmonds P, Adams MJ, Benkö M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 2017;15:161-168. doi: 10.1038/nrmicro.2016.177.

2. Gregory AC, Zayed AA, Concieção-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell. 2019;177(5):1109-1123. doi: 10.1016/j.cell.2019.03.040.

3. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiology Reviews. 2016;40(2):258-272. doi: 10.1093/femsre/fuv048.

4. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLoS Biol. 2023;21(4):e3002083. doi: 10.1371/journal.pbio.3002083.

5. Dion MB, Plante P-L, Zufferey E, Shah SA, Corbeil J, Moineau S. Streamlining CRISPR spacer-based bacterial host prediction to decipher the viral dark matter. Nucleic Acid Research. 2021;49(6):3127-3138. doi. 10.1093/nar/gkab133.

6. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves predictions of hosts from metagenomically-derived viral sequences. Nucleic Acids Research. 2017;45(1):39-53. doi: 10.1093/nar/gkw1002.

7. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017;33(19):3113-3114. doi: 10.1093/bioinformatics/btx383.

8. Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, Wu A, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biol. 2021;19(5). doi: 10.1186/s12915-020-00938-6.

9. Wommack KE, Colwell RR. Virioplankton: Viruses in Aquatic Ecosystems. Microbiology and Molecular Biology Reviews. 2000;64(1):69-114. doi: 10.1128/MMBR.64.1.69-114.2000.

10. Winter C, Bouvier T, Weinbauer MG, Thingstad TF. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "Killing the Winner" hypothesis revisited. Microbiology and Molecular Biology Reviews. 2010;74:42-57. doi: 10.1128/MMBR.00034-09.

11. Coutinho FH, Zaragoza-Salas A, López-Pérez M, Barylski J, Zielezinski A, Dutilh BE, et al. RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. CellPress. 2021;2(7). doi: 10.1016/j.patter.2021.100274.

12. Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, et al. HostPhinder: A Phage Host Prediction Tool. Viruses. 2016;8(5):116. doi: 10.3390/v8050116.

13. Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, et al. A network-based integrated framework for predicting virus–prokaryote interactions. NAR Genomics and Bioinformatics. 2020;2(2):lqaa044. doi: 10.1093/nargab/lqaa044.

14. Zhou F, Gan R, Zhang F, Ren C, Yu L, Si Y, et al. PHISDetector: A tool to detect diverse in silico phage-host interaction signals for virome studies. Genomics, Proteomics & Bioinformatics. 2022;20(3):508-523. doi: 10.1016/j.gpb.2022.02.003.

15. Altschul SF, Gish W, Miller w, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215(3):403-410. doi: 10.1016/S0022-2836(05)80360-2.s

16. Medina-Chávez N-O, Vildaomat-Jasso M, Zarza E, Islas-Robles A, Valdivia-Anistro J, Thalasso-Siret F, et al. A transiently hypersaline microbial mat harbors a diverse

and stable archaeal community in the Cuatro Cienegas Basin, Mexico. Astrobiology. 2023;8. doi: 10.1089/ast.2021.0047.

17. Espinosa-Asuar L, Monroy-Guzmán C, Madrigal-Trejo D, Navarro-Miranda M, Sánchez-Pérez J, Buenrostro-Muñoz J, et al. Diversity of an uncommon elastic hypersaline microbial mat along a small scale transect. PeerJ. 2022;10:e13579. doi: 10.7717/peerj.13579.
18. Cisneros-Martínez AM, Eguiarte LE, Souza V. Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat. Microbial Genomics. 2023;9(7). doi: 10.1099/mgen.0.001063.
19. Biswas A, Staals RHJ, Morales SE, Fineran PC, Brown CM. CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics 2016;17:356. 10.1186/s12864-016-2627-0.
20. Purdy KJ. Nucleic acid recovery from complex environmental samples. Methods Enzymol. 2005;397:271–292. doi: 10.1016/S0076-6879(05)97016-X.
21. De Anda V, Zapata-Peñasco I, Blaz J, Poot-Hernández AC, Contreras-Moreira B, González-Laffitte M, et al. Understanding the mechanisms behind the response to environmental perturbation in microbial mats: A metagenomic-network based approach. Frontiers in Microbiology. 2018;9:2606. doi: 10.3389/fmicb.2018.02606.
22. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Version 0.11.9 [software]. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014;30(15):2114-2120. doi: 10.1093/bioinformatics/btu170.

24. Antipov D, Raiko M, Lapidus A, Pevzner PA. METAVIRALSPADES: assembly of viruses from metagenomic data. Bioinformatics. 2020;36(14):4126-4129. doi: 10.1093/bioinformatics/btaa490.

25. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16(6):276-277. doi: 10.1016/s0168-9525(00)02024-2.

26. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. doi: 0.1186/1471-2105-11-119.

27. Eddy SR. Accelerated profile HMM searches. PLoS Comp. Biol. 2011;7:e1002195. doi: 10.1371/journal.pcbi.1002195.

28. Nurk S, Meleshko D, Korobeynikov A, Pevzner P. metaSPAdes: a new versatile metagenomic assembler. Genome Research. 2017;27(5):824-834. doi: 10.1101/gr.213959.116.

29. Wu Y, Simmons B, Singer S. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2015;32(4):605-607. doi: 10.1093/bioinformatics/btv638.

30. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7. doi: 10.7717/peerj.7359.

31. Song W, Thomas T. Binning_refiner: improving genome bins through the combination of different binning programs. Bioinformatics. 2017;33(12):1873-1875. doi: 10.1093/bioinformatics/btx086.

32. Parks D, Imelfort M, Skennerton C, Hugenholtz P, Tyson G. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015;25(7):1043-1055. doi: 10.1101/gr.186072.114.

33. Chaumeil P, Mussig A, Hugenholtz P, Parks D. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36(6):1925-1927. doi: 10.1093/bioinformatics/btz848.

34. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ. CRISPRCASTYPER: Automated Identification, annotation, and classification of CRISPR-Cas Loci. The CRISPR Journal. 2020;3(6):462–469. doi: 10.1089/crispr.2020.0059.

35. Madrigal-Trejo D, Sánchez-Pérez J, Espinosa-Asuar L, Valdivia-Anistro JA, Eguiarte LE, Souza V. A metagenomic time-series approach to assess the ecological stability of microbial mats in a seasonally fluctuating environment. Microbial Ecology. doi: 10.1007/s00248-023-02231-9.

36. Challacombe JF, Majid S, Deole R, Brettin TS, Bruce D, Delano SF, et al. Complete genome sequence of *Halorhodospira halophila* SL1. Stand. Genomic Sci. 2013;8(2):206-214. doi: 10.4056/sigs.3677284.

37. Sorokin DY, Tourova TP, Galinski EA, Muyzer G, Kuenen JG. *Thiohalorhabdus denitrificans* gen. nov., sp. nov., an extremely halophilic, sulfur-oxidizing, deep-lineage gammaproteobacterium from hypersaline habitsts. Int J Sys Evol Microbiol. 2008;58(Pt 12):2890-7. doi: 10.1099/ijs.0.2008/000166-0.

38. Kumar PA, Srinivas TNR, Sasikala Ch, Ramana ChV. *Halochromatium roseum* sp. nov., a non-motile phototrophic gammaproteobacterium with gas vesicles, and

emended description of the genus *Halochromatium*. Int J Sys Evol Microbiol. 2007;57(9):2110-2113. doi: 10.1099/ijs.0.65034-0.

39. Hirsch P, Hoffmann B. *Dichotomicrobium thermohalophilum*, gen. nov., spec. nov., budding prosthecate bacteria from the Solar Lake (Sinai) and some related strains. System. Appl. Microbiol. 1989;11:291-301. doi: 10.1016/S0723-2020(89)80027-X.
40. Nathani NM, Dave KJ, Vatsa PP, Mahajan MS, Sharma P, Mootapally C. 309 metagenome assembled microbial genomes from deep sediment samples in the Gulfs of Kathiawar Peninsula. Sci Data. 2021;8:194. doi: 10.1038/s41597-021-00957-0.
41. Breuker A, Köweker G, Blazejak A, Schippers. The deep biosphere in terrestrial sediments in the Chesapeake Bay area, Virginia, USA. Front Microbiol. 2011;2. doi: 10.3389/fmicb.2011.00156.
42. McGonigle JM, Bernau JA, Bowen BB, Brazelton WJ. Metabolic potential of microbial communities in the hypersaline sediments of the Bonneville Salt Flats. mSystems. 2022;7(6). doi: 10.1128/msystems.00846-22.
43. Sorokin DY, Gumerov VM, Rakitin AL, Beletsky AV, Damsté JSS, Muyzer G, et al. Genome analysis of *Chitinivibrio alkaliphilus* gen. nov., sp. nov., a novel extremely haloalkaliphilic anaerobic chitinolytic bacterium from the candidate phylum Termite Group 3. Environ Microbiol. 2014;16(6):1549-65. doi: 10.1111/1462-2920.12284.
44. Kelly DP, Wood AP. *Halothiobacillaceae* fam. nov. *Bergey's Manual of Systematics of Archaea and Bacteria*. 2015:1-2. doi:10.1002/9781118960608.fbm00221.
45. Sorokin SY, Mosier D, Zorz JK, Dong X, Strous M. *Wenzhouxiangella* strain AB-CW3, a proteolytic bacterium from hypersaline soda lakes that preys on cells of Gram-positive bacteria. 2020;11. doi: 10.3389/fmicb.2020.597686.

46. Xia J, Zhao J-X, Sang J, Chen G-J, Du Z-J. *Halofilum ochraceum* gen. nov., sp. nov., a gammaproteobacterium isolated from a marine solar saltern. Int J Sys Evol Microbiol. 2017,67(4):932-938. doi: 10.1099/ijsem.0.001718.

47. Fukunaga Y, Kurahashi M, Sakiyama Y, Ohuchi M, Yokota A, Harayama S. *Phycisohaera mikurensis* gen. nov., sp. nov., isolated from a marine alga, and proposal of *Phycisphaeraceae* fam. nov., *Phycisphaerales* ord. nov. and *Phycisphaerae* classis nov. in the phylum *Planctomycetes*. J Gen Appl Microbiol. 2009;55(4):267-75. doi: 10.2323/jgam.55.267.

48. Alfredsson GA, Kristjansson JK, Hjrleifsdottir S; Stetter KO. *Rhodothermus marinus*, gen. nov., sp. nov., a thermophilic, halophilic bacterium from submarine hot springs in Iceland. Microbiology. 1988;134(2):299-306. doi:10.1099/00221287-134-2-299.

49. Sorokin DY, Khijniak TV, Galinski EA, Kublanov IV. *Natronotalea proteinilytica* gen. nov., sp. nov. and *Longimonas haloalkaliphila* sp. nov., extremely haloalkaliphilic members of the phylum *Rhodothermaeota* from hypersaline alkaline lakes. Int J Sys Evol Microbiol. 2017;67(10):4161-4167. doi:10.1099/ijsem.0.002272.

50. Miranda-Tello E, Fardeau M-L, Thomas P, Ramirez F, Casalot L, Cayol J-L. *Petrotoga Mexicana* sp. nov., a novel thermophilic, anaerobic and xylanolytic bacterium isolated from an oil-producing well in the Gulf of Mexico. Int J Syst Evol Microbiol. 2004;54(Pt 1):169-174. doi: 10.1099/ijs.0.02702-0.

51. Oren A. The order *Halanaerobiales*, and the families *Halanaerobiaceae* and *Halobacteroidaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackerbrandt E,

Thompson F, editors. The Prokaryotes. Berlin, Heidelberg: Springer; 2014. doi: 10.1007/978-3-642-30120-9_218.

52. Íñiguez-Martínez AM, Cardoso-Martínez F, de la Rosa J, Cueto M, Díaz-Marrero A, Darias J, et al. Compound isolated from *Salinispora Arenicola* of the Gulf of California, México. Revista de Biología Marina y Oceanografía. 2016;51(1):161-170. doi: 10.4067/S0718-19572016000100015.

53. Marietou A. Chapter two – Sulfate reducing microorganisms in high temperature oil reservoirs. Advances in Applied Microbiology. 2021;116:99-131. doi: 10.1016/bs.aambs.2021.03.004.

54. Albuquerque L, da Costa MS. The family *Thermaceae*. In: Rosenberg E, DeLong EF, Lory S, Stackerbrandt E, Thompson F, editors. The Prokaryotes. Berlin, Heridelberg: Springer; 2014. doi: 10.1007/978-3-642-38954-2_128.

55. Entcheva-Dimitrov P, Spormann AM. Dynamics and control of biofilms of the oligotrophic bacterium *Caulobacter crescentus*. J Bacteriol. 2004;186(24):8254-8266. doi: 10.1128/JB.186.24.8254-8266.2004.

56. Kamruzzaman M, Iredell JR. CRISPR-Cas System in Antibiotic Resistance Plasmids in *Klebsiella pneumoniae*. Front. Microbiol. 2020;10:2934. doi: 10.3389/fmicb.2019.02934.

57. Souza V, Espinosa-Asuar L, Escalante AE, Eguiarte LE, Farmer J, Forney L, et al. An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. Proc Natl Acad Sci. 2006;103:6565–6570. doi: 10.1073/pnas.0601434103.

58. Souza V, Siefert JL, Escalante AE, Elser JJ, Eguiarte LE. The Cuatro Ciénegas Basin in Coahuila, Mexico: an astrobiological Precambrian park. Astrobiology. 2012;12:641–647. doi: 10.1089/ast.2011.0675.

59. Wolaver BD, Crossey LJ, Karlstrom KE, Banner JL, Cardenas MB, Gutiérrez-Ojeda C, et al. Identifying origins of and pathways for spring waters in a semiarid basin using He, Sr, and C isotopes: Cuatrocienegas Basin, Mexico. Geosphere. 2013;9:113–125. doi: 10.1130/GES00849.1.s

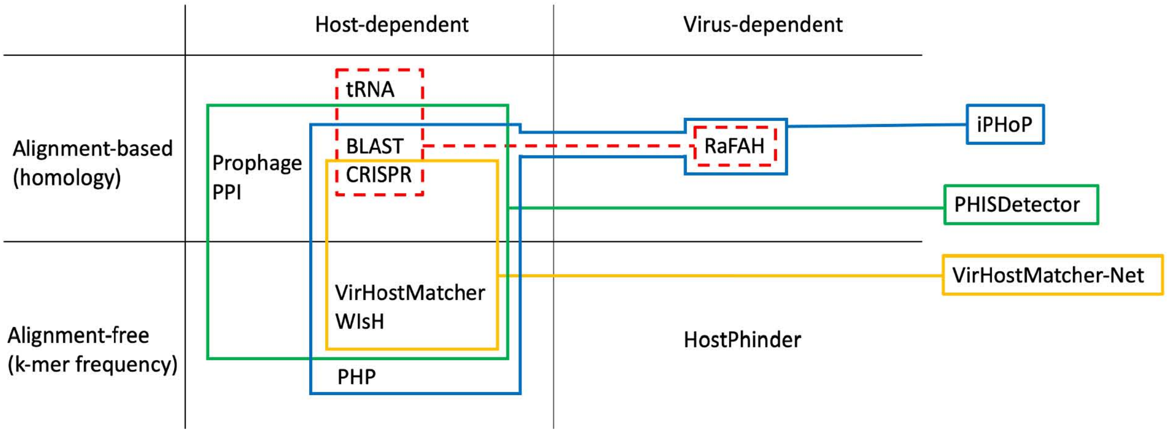
Supporting information

S1 File. Lists of NCBI complete bacterial virus genomes and RefSeq complete bacterial genomes used for benchmarking virus-host prediction tools. The file contains six sheets—the first one lists bacteriophage genomes. The second is the Virus-Host DB table. The third one is the RefSeq release catalog with complete bacterial genomes—the fourth lists the virus-host pairs, including their accessions, used for benchmarking. The fifth sheet is the reference accession-genus list against which each prediction was compared. The sixth and final sheet contains the prediction results of each tool, using the parameters on which they perform the best.

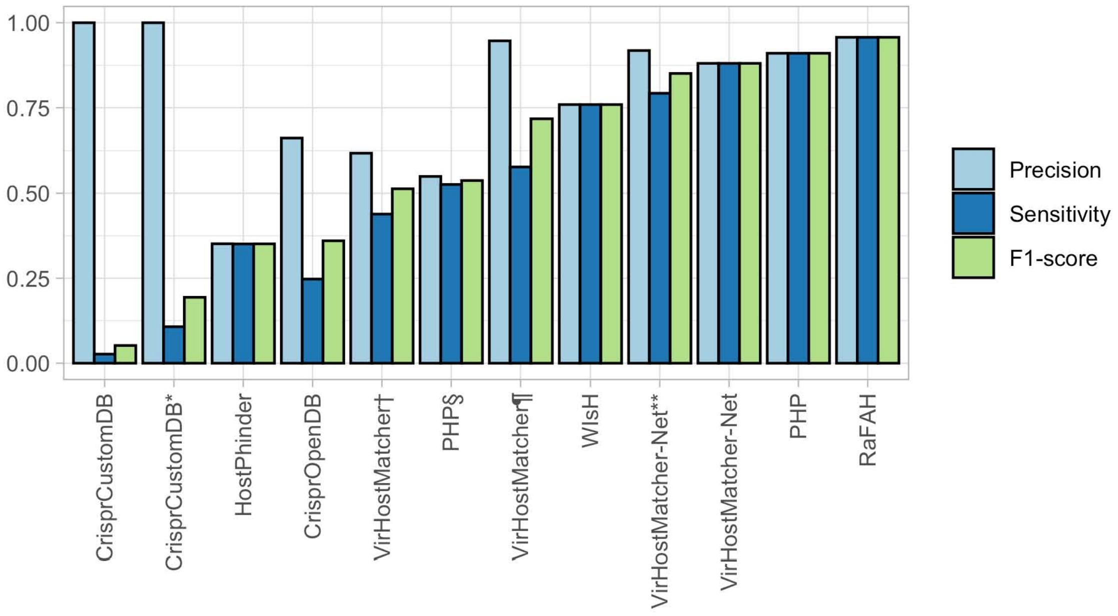
S2 File. NCBI complete bacterial virus genomes and RefSeq complete bacterial genomes sequence files used for benchmarking virus-host prediction tools. It includes viral and bacterial genomes in fasta format. It also contains files with predicted CRISPR arrays and spacers.

S3 File. Files used for testing virus-host prediction tools on metagenomic data from Archaean Domes, Cuatro Ciénegas Basin, Mexico. It includes spacers needed to run predictions with BLAST and CrisprCustomDB and the host k-mer file needed to run predictions with PHP.

S4 File. Host prediction results on mVCs from Archaean Domes, Cuatro Ciénegas Basin, Mexico. It includes a table with contig information and host predictions and a Venn diagram showing the number of shared predictions by the different tools. Contigs highlighted with beige backgrounds will likely represent the same virus according to their protein domain content and host prediction. A red line delimits unreliable predictions from predictions supported by only one method but with consistency between the predicted host biology and the virus source environment. Above the yellow line, predictions are supported by two methods. Above the green line, predictions are supported by three methods.



Host-dependent
Virus-dependent
Alignment-based
(homology)
Alignment-free
(k-mer frequency)
tRNA
BLAST
CRISPR
Prophage
PPI
VirHostMatcher
WIsH
PHP
RaFAH
HostPhinder
iPHoP
PHISDetector
VirHostMatcher-Net



1.00
0.75
0.50
0.25
0.00
CrisprCustomDB
CrisprCustomDB*
HostPhinder
CrisprOpenDB
VirHostMatcher†
PHP§
VirHostMatcher¶
WIsH
VirHostMatcher-Net**
VirHostMatcher-Net
PHP
RaFAH
Precision
Sensitivity
F1-score

Artículo anexo: Ancient gene duplications in RNA viruses revealed by protein tertiary structure comparisons

VIRUS EVOLUTION

Virus Evolution, 2021, 7(1): veab019

doi: 10.1093/ve/veab019
Research Article

Ancient gene duplications in RNA viruses revealed by protein tertiary structure comparisons

Alejandro Miguel Cisneros-Martínez,[1] Arturo Becerra,[1] and Antonio Lazcano[1,2,*]

[1]Facultad de Ciencias, Universidad Nacional Autónoma de México, Mexico City, Mexico and [2]El Colegio Nacional, Donceles 104, Centro Histórico, Mexico City, Mexico

*Corresponding author: E-mail: alar@ciencias.unam.mx

Abstract

To date only a handful of duplicated genes have been described in RNA viruses. This shortage can be attributed to different factors, including the RNA viruses with high mutation rate that would make a large genome more prone to acquire deleterious mutations. This may explain why sequence-based approaches have only found duplications in their most recent evolutionary history. To detect earlier duplications, we performed protein tertiary structure comparisons for every RNA virus family represented in the Protein Data Bank. We present a list of thirty pairs of possible paralogs with <30 per cent sequence identity. It is argued that these pairs are the outcome of six duplication events. These include the α and β subunits of the fungal toxin KP6 present in the dsRNA *Ustilago maydis virus* (family *Totiviridae*), the SARS-CoV (*Coronaviridae*) nsp3 domains SUD-N, SUD-M and X-domain, the *Picornavirales* (families *Picornaviridae, Dicistroviridae, Iflaviridae and Secoviridae*) capsid proteins VP1, VP2 and VP3, and the *Enterovirus* (family *Picornaviridae*) 3C and 2A cysteine-proteases. Protein tertiary structure comparisons may reveal more duplication events as more three-dimensional protein structures are determined and suggests that, although still rare, gene duplications may be more frequent in RNA viruses than previously thought.

Keywords: gene duplications; RNA viruses.

1. Introduction

Many hypotheses on the evolutionary importance and the mechanisms of gene duplications were already established by cytologists and cytogeneticists since the first decades of the twentieth century (Taylor and Raes 2004). However, it was not until the publication of *Evolution by gene duplication* by Susumu Ohno (1970), when the idea of gene duplication as a major evolutionary force became widely acknowledged. During the following decades, with the advent of DNA sequencing techniques, a wealth of accumulated evidence contributed considerably to our understanding of gene duplications, allowing for the refinement of models that describe its mechanisms and underlying its evolutionary relevance (Taylor and Raes 2004). The rationale for understanding the evolutionary significance of gene duplications lies within the notion that evolution cannot proceed solely through point mutations because any mutation that alters the function of a coding gene would be deleterious. The solution to this conundrum is provided by Ohno (1970): 'Only the cistron which became redundant was able to escape from the relentless pressure of natural selection, and by escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus'. The evolutionary significance of gene duplications is highlighted by the relatively high frequency at which they occur in the three major domains of life (17–44% in bacteria, ~30% in archaea and 30–65% in eukarya) (Zhang 2003). In fact, the identification of ancient gene duplications that appear to have happened before the divergence of the three domains of life (Becerra et al. 2007) suggests that it has been one of the most important mechanisms for increasing the size and complexity of genomes since the early stages of cell evolution (Lazcano 1995).

Previous studies have shown that gene duplications have been a relatively frequent event in dsDNA viral genome evolution (Shackelton and Holmes 2004). Many examples of gene duplications are known in *Adenoviridae* (Davison et al. 2003), *Herpesviridae* (McGeoch and Davison 1999) and *Poxviridae* (Hughes and Friedman 2005). A search on 201 dsDNA viruses found gene duplications in 42.3 per cent of its genomes. The 1,874 identified paralogs were distributed in 612 protein families with two to sixty-one members (Gao et al. 2017). Additionally, a positive correlation was found between paralog number and genome size, which can reach up to 2,473 kbp in *Pandoravirus salinus*. In sharp contrast, Simon-Loriere and Holmes (2013) detected gene duplications only in 19 out of 1,198 (1.6%) RNA viruses analysed. The twenty paralogs were distributed in eight protein families composed of two to three members (Table 1). These twenty pairs are likely to represent nine duplication events that include four cases in ssRNA(+) viruses: 1) the coat protein (CP) and the minor CP (CPm) in the family *Closteroviridae* (Boyko et al. 1992; Kreuze et al. 2002; Tzanetakis et al. 2005; Tzanetakis and Martin 2007; Simon-Loriere and Holmes 2013), as well as a tandem duplication of CPm in the *Grapevine leafroll-associated virus* 1 (Fazeli and Rezaian 2000); 2) p25 and p26 proteins encoded by the third and fifth segments, respectively, in the family *Benyviridae* (Simon-Loriere and Holmes 2013); and 3) a tandem duplication of the genome-linked protein VPg in the *Foot-and-mouth disease virus* of the family *Picornaviridae* (Forss and Schaller 1982). In ssRNA(-), the two cases were found in the family *Rhabdoviridae*: 1) the G and Gns glycoproteins in some viruses of the genera *Ephemerovirus* (Walker et al. 1992; Blasdell et al. 2012) and *Hapavirus* (Gubala et al. 2010); and 2) U1 and U2 of unknown function (Simon-Loriere and Holmes 2013). Finally, in ssRNA(RT), three cases were found in the family *Retroviridae*: 1) orfA and orfB of *Walleye epidermal hyperplasia virus 2* (LaPierre et al. 1999); 2) orf1 and orf2 of *Xenopus laevis endogenous retrovirus* (Kambol et al. 2003); and 3) vpr and vpx in *Human immunodeficiency virus* 2 and *Simian immunodeficiency virus—mnd* 2 (Tristem et al. 1990). As of today, no gene duplication events have been reported in dsRNA viruses.

Detection of gene duplications in RNA viral genomes is complicated for a number of reasons. For instance, the number of paralogs is known to be positively correlated with the genome size (Gevers et al. 2004) and, with the exception of coronaviruses, RNA viruses tend to have smaller genomes (from ~2 to ~33 kbp) compared with dsDNA viruses (from ~5 to ~2,500 kbp) (Campillo-Balderas et al. 2015). The genome sizes of RNA viruses may be limited by the high error rate of RNA replicases (Reanney 1982; Holmes 2009). As described by Eigen (1971), nucleic acids need a minimum replication fidelity to preserve the genetic information, where a higher fidelity allows a higher information content. This implies that the amount of genetic information is limited by the precision of the copying process. If the genome grows beyond the error, threshold deleterious mutations would quickly appear (Eigen 1971; Maynard Smith and Szathmáry 1995). In fact, an inverse relationship between mutation rate and genome size has been observed from viroids to eukaryotes, in which RNA viruses appear as the second biological entities with the highest mutation rates and the shortest genomes (Gago et al. 2009; Holmes 2011). Paradoxically, to evolve an accurate replication machinery requires more coding capabilities and thus a larger genome. This so-called Eigen's paradox (Maynard Smith and Szathmáry 1995) might imply that most RNA virus genomes are irrevocably limited to remain small. As can be inferred from Sol Spiegelman's *in vitro* RNA replication and evolution experiment from 1970 (Maynard Smith and Szathmáry 1995), replication efficiency is another pressure that selects for smaller genomes, which could also affect RNA viruses that benefit from a faster replication in a context of competition with the host and other viruses for cellular resources.

Other factors that could underline the RNA viruses genome size restrictions include the shape and size of the capsid and the impossibility of unwinding large dsRNA structures formed during the replication in viruses lacking a helicase domain (Reanney 1982; Holmes, 2009). Finally, and as expected, in a directed evolution experiment that tested the stability and fitness effect of different duplicated genes in artificial constructs derived from the plant-infecting ssRNA(+) *Tobacco etch virus* (family *Potyviridae*), Willemsen et al. (2016) observed a fitness reduction and the deletion of the duplicated gene. As a further explanation to the deleterious effect of gene duplications, they suggested that the correct processing of the polyprotein and a greater cellular resource requirement to express a larger genome could be important factors contributing to the constraints on RNA virus genome size. Some of these issues may also apply to ssDNA viruses which, from an evolutionary perspective, have been thought to behave similarly to RNA viruses, showing small genome sizes and little gene duplication (Boyko et al. 1992; Holmes 2009).

The current evidence of gene duplications in RNA viruses comes mainly from protein primary structure which, given the previously mentioned high mutation rate, can only recognize the most recent duplications (Simon-Loriere and Holmes 2013). As argued here, protein tertiary structure comparisons can broaden the known universe of paralogous proteins in RNA viruses. Our results suggest that in fact gene duplications might be more stable in RNA genomes than previously thought.

2. Methods

2.1 Data selection

On 17 July 2020, we performed an advanced search on RCSB Protein Data Bank (PDB) (www.rcsb.org) (Berman et al. 2000) based on the following criteria:

Table 1. Gene duplications are much more frequent in dsDNA compared with RNA viruses.

	dsDNA viruses (Gao et al. 2017)	RNA viruses (Simon-Loriere and Holmes 2013)
Viruses with duplicated genes	85/201 (42.3%)	19/1198 (1.6%)
Number of paralogous pairs	1874	20
Number of paralogous families	612	8
Family size	2 to 61 members	2 to 3 members

The table summarizes the results described by Gao et al. (2017) and Simon-Loriere and Holmes (2013), respectively.

- Source Organism Taxonomy Name equals Riboviria
- Polymer Entity Distinct Taxonomy Count = 1
- Experimental Method equals X-RAY DIFFRACTION
- Resolution ≤ 3Å
- Polymer Entity Type equals Protein
- Polymer Entity Sequence Length ≥ 80
- Polymer Entity Mutation Count = 0

The search resulted in a table (Supplementary data S1) describing different features (such as Entity ID, Number of Entities (Protein), PDB ID, Source Organism, Taxonomy ID, Macromolecule Name, Resolution Å, R Work, Deposition Date, Structure Title, Chain Length, Number of Polymer Residues, Entity Polymer Type, Structure Keywords, PubMed Central ID, PubMed ID, DOI) of 4,049 protein entities. To select representatrive structures, the information was sorted on the basis of:

- PDB in alphabetical order
- Deposition date (most recent)
- Quality factor (highest)

$$o \frac{1}{Resolution\ \text{Å} - \text{R}\ Work}$$

- Macromolecule name in alphabetical order
- Source organism in alphabetical order

The manual selection resulted in 1,112 representative entities corresponding to 961 PDB IDs (Supplementary data S2). The corresponding sequences were downloaded from RCSB PDB and further redundancy was filtered with a three-step iterative hierarchy clustering with CD-HIT (90% and 60% sequence identity) and PSI-CD-HIT (30% sequence identity) (Li and Godzik 2006). The clustering resulted in 305 protein entities corresponding to 297 PDB IDs (Supplementary data S3 and S4), which were downloaded from RCSB PDB and parsed with the Perl module ParsePDB.pm (Bulheller and Hirst 2009) to retrieve only the corresponding chains. Taxonomic annotation based on the NCBI taxonomy (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/ accessed July 2020) revealed that the most frequent entries in the dataset corresponded to ssRNA(+), followed by ssRNA(-), dsRNA and ssRNA(RT) viral families and two ssRNA(+) satellite viruses.

2.2 Tertiary structure alignments

Structural comparisons were carried out for each viral family. Families with only one representative structure (*Picobirnaviridae*, *Alphaflexiviridae*, *Alphatetraviridae*, *Mesoniviridae*, *Permutotetraviridae*, *Potyviridae*, *Tospoviridae*) and viruses with no family assigned (satellite viruses) were excluded from the analysis. A total of 2,406 tertiary structure alignments (Supplementary data S5) were automatically conducted via the FATCAT rigid algorithm (Ye and Godzik 2003). Structural similarity was evaluated with a modification of the structural alignment score (SAS) (Subbiah et al 1993) that takes into account the alignment coverage of each structure defined as:

$$Lsas = \frac{100\text{RMSD}(L1 + L2)}{2Naln^2}$$

Where Lsas stands for length-weighted SAS, RMSD is the root mean square deviation between α-carbon atoms, L1 and L2 correspond to the lengths of the superposed structures and Naln is the number of aligned residues. 257 alignments with Lsas < 10 (Supplementary data S6) were manually analyzed to evaluate the homology type between pairs. Likely paralogs were defined with Lsas < 5. Structural alignments were visualized with UCSF Chimera (Pettersen et al. 2004).

2.3 Structure similarity trees

Structural models 2acf-A (X-domain), 1b35-C (VP3), 3q3y-A (3C) and both chains in 4gvb (KP6α and KP6β) were queried for a PDB search within the DALI server (http://ekhidna2.biocenter.helsinki.fi/dali/ accessed January 2021) (Holm 2020). Models selection for the structure similarity tree analyses was performed as follows: 1) for the KP6 subunits only the shared hits with Z ≥ 4 on the PDB25 report; 2) for the X-domain hits with Z ≥ 12 on the PDB90 report; 3) for VP3 hits with Z ≥ 7 on the PDB25 report (*Secoviridae* models 1a6c, 1bmv and 7chk were excluded to avoid long branch attraction artifacts); and 4) for the 3C protease non-viral hits with Z ≥ 11 on the PDB25 report plus viral hits with Z ≥ 11 on the PDB90 report (Supplementary data S8). In each case, protein and species redundancies were omitted. The models were edited with UCSF Chimera to retain only the corresponding chains. For the proteases, only the carboxy terminal domains were used. Multiple structure alignments were performed with the STAMP algorithm (Russell and Barton 1992) within the MultiSeq tool (Roberts et al. 2006) in VMD 1.9.3 (Humphrey et al. 1996) with default parameters for the KP6 subunits and the 3C protease and its respective relatives, and with scanscore = 0 for VP3 and its related structures. The X-domain-related structures were compared through pairwise against all alignments with the MatchMaker tool within Chimera using a BLOSUM30 matrix and the Smith-Waterman algorithm. Structure similarity was assessed with the Match → Align tool in Chimera from which RMSD and number of aligned residues were retrieved to compute SAS (100RMSD/Naln) as a geometric distance measure. The resulting distance matrices were introduced into the FITCH algorithm (Fitch and Margoliash 1967) within the PHYLIP 3.695 package (Felsenstein 1989) for tree construction with global branch-swapping rearrangements and the jumble option to randomize 100 times the input order. For the capsid proteins, the tree was rooted on *Solemoviridae* and *Tombusviridae* single jelly roll capsid proteins as outgroup. For the rest, the root was placed using the MAD method (Tria et al. 2017). Finally, tree visualization was made with FigTree 1.4.2 (Rambaut 2014).

3. Results

A total of 30 pairs of likely paralogous proteins was found (Supplementary data S7). Table 2 shows 12 representative pairs with the lowest Lsas. As argued below, it is possible that these cases represent six duplication events: 1) one that led to the α and β subunits of the *Ustilago maydis virus* (UmV) (family *Totiviridae*) KP6 toxin, which is potentially the first confirmed case of gene duplication in a dsRNA virus; 2) the SARS-Unique domain (SUD) N and M domains found in sarbecoviruses (family *Coronaviridae*) that may come from the more widely distributed coronavirus X-domain; 3) a duplication event that probably gave rise to the chymotrypsin-related 2A cysteine protease in the genus *Enterovirus* (family *Picornaviridae*) from the greater distributed 3C cysteine protease; and 4) VP1, VP2 and VP3, that probably originated after two duplication events during the dawn of the order *Picornavirales* (Liljas et al. 2002).

Table 2. Likely paralogs detected by protein tertiary structure comparisons.

Genome type	Order	Family	Pair	RMSD	Naln	Lsas
dsRNA	*Ghabrivirales*	*Totiviride*	KP6α and KP6β	1.52	62	2.9854
ssRNA(+)	*Nidovirales*	*Coronaviridae*	SUD-M and X-domain	2.82	123	3.9982
	Picornavirales	*Dicistroviridae*	VP1 and VP2	3.14	173	2.6596
			VP1 and VP3	3.11	208	2.0164
			VP2 and VP3	3.13	208	2.0112
		Iflaviridae	VP1 and VP2	3.02	173	2.5831
			VP1 and VP3	3.17	206	2.4987
			VP2 and VP3	3.07	182	3.1373
		Picornaviridae	VP1 and VP2	3.19	165	2.8883
			VP1 and VP3	3.15	179	2.3398
			VP2 and VP3	3.08	167	2.8438
			3C and 2 A	2.75	140	2.3151

The table shows 12 representative pairs with the lowest Lsas. Detailed information on the 30 protein pairs and PDB IDs is available in Supplementary data S7.

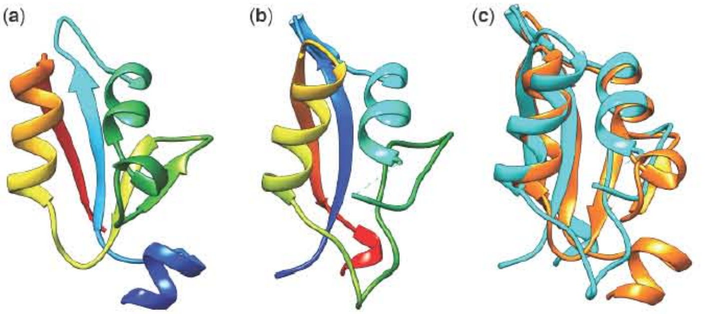


Figure 1. Tertiary structure of KP6α (a) and KP6β (b). Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (c) 3D alignment between KP6α (orange) and KP6β (cyan) as made by FATCAT rigid. PDB ID 4GVB.

3.1 *Totiviridae* KP6α-KP6β

As described elsewhere (Allen et al. 2013), KP6 is a viral toxin present in UmV. This virus is found only in fungi containing resistance genes that allow them to compete with other strains not resistant to the toxin. Thus, UmV acts as a symbiont which is only transmitted from one cell to another through mitosis or meiosis. KP6 is a heterodimer whose subunits are encoded on a single satellite dsRNA. Both subunits are translated as a single polypeptide that undergoes protease cleavage on a 31 amino acid linker between the former amino (KP6α) and carboxy (KP6β) terminal domains. KP6α and KP6β are 77 and 74 residues long, respectively, both of which fold into a α/β sandwich structure consisting of a four-stranded antiparallel β-sheet and a pair of antiparallel α-helices. The major differences between these structures are the presence of an extra N-terminal helix in KP6α, and longer α2-β2 and β3-α3 loops in KP6β (Fig. 1).

Although no clear statement regarding the paralogous relationship between the genes was made, a 3D alignment of the two KP6 subunits had been reported by Allen et al. (2013) with very similar results as those reported here with FATCAT rigid. Interestingly, upon structural database search, the only similar proteins to KP6α and KP6β are individual domains within larger cellular proteins, with KP6 being the only protein showing this kind of heterodimer (Allen et al. 2013). This suggests that the virus did not acquire an already duplicated protein, but that the gene encoding it underwent a duplication event in the virus after the acquisition of a single domain.

3.2 *Sarbecovirus* Sud and X-domain

SARS-CoV Nsp3 is translated as a polyprotein that contains an acidic domain, an X-domain, the SUD, a papain-like cysteine protease domain and other domains including a transmembrane region. The X-domain is a homodimeric phosphatase that removes the 1′ phosphate group of ADP-ribose-1′-phosphate (ADRP). Its structure is mainly defined by seven central β strands surrounded by six α helices (three on each side of the sheet), conforming a three-layered α/β/α topology as seen in proteins belonging to the Macro-H2A fold. The innermost five β strands are parallel while the outermost two strands are antiparallel (Saikatendu et al. 2005). SUD is a domain present only in SARS-CoV and closely related sarbecoviruses. It is known to be endowed with two macrodomains, N and M, similar to ADRP which is also present in other coronaviruses and even in viruses belonging to different families. However, SUD domains lack phosphatase activity, and have been shown to bind to oligonucleotides forming G-quadruplex secondary structures. The structure of SUD-N consists of six β strands and four α helices, while SUD-M is made of six β strands and five α helices. In both domains, the β sheet has five parallel strands and only one antiparallel β3 strand (Tan et al. 2009).

Although the homologous relations between SUD-N, SUD-M and the X-domain was not discussed by Tan et al. (2009), they had in fact described their structural similarity. They superimposed SUD-N and SUD-M with an RMSD of 3.3 Å, and found a conserved Leu-Glu-Glu-Ala motif at the N-terminal end of helix α4. We have confirmed the structural similarity between SUD-M and the X-domain (Fig. 2). Tan et al. observed better RMSD values between SUD-M and X-domain (2.3 Å) than between SUD-N and X-domain (2.7 Å), which is consistent with our results. Given the taxonomic distribution of these domains, the adjacent position of the three domains in the Nsp3 polyprotein and the clear structural similarity, we posit that SUD-N and SUD-M are likely paralogs that resulted from two duplication events that started with the duplication of the X-domain.

3.3 *Picornavirales* capsid proteins VP1, VP2 and VP3

The order *Picornavirales* comprises families such as *Dicistroviridae*, *Iflaviridae*, *Marnaviridae*, *Picornaviridae* and *Secoviridae*. In most viruses belonging to the family *Picornaviridae*, the capsid genes are translated into a single polyprotein which is proteolytically cleaved into VP0, VP3 and VP1. VP0 is then self-cleaved into VP4 and VP2, except in the genera *Kobuvirus* and *Parechovirus*, in which the equivalent of VP4 remains as a N-terminal extension of VP2 (Sabin et al. 2016; Kalynych et al. 2016a). In other families, such as *Dicistroviridae* and *Iflaviridae*, VP4 is cleaved from the N-terminal region of VP3 (Liljas et al. 2002; Kalynych et al. 2016b). In viruses of the family *Secoviridae* there is no equivalent to VP4, and in some cases the polyprotein is partially cleaved into a large and small subunit made of two and one domains, respectively (e.g. *Comovirus*), or not cleaved at all (e.g. *Nepovirus*). VP1, VP2 and VP3, or its equivalent domains (A, C and B) are the main building blocks of the *Picornavirales* capsids. These proteins are jelly-roll β barrels of approximately 250 residues long made of eight anti-parallel strands (B-I). Sixty copies of each domain assembly to form a T = p3 icosahedral capsid with a diameter of approximately 30 nm (Rossmann and Johnson 1989).

As discussed in qualitative terms by Chandrasekar and Johnson (1998) and Liljas et al. (2002), VP1, VP2 and VP3 tertiary structures are remarkably similar. As shown in Fig. 3, this similarity is particularly clear after visual inspection of the capsid proteins of the family *Dicistroviridae*, in which VP1, VP2 and VP3 do not have large insertions and the N-terminal arm conformation is conserved. Given that VP1, VP2 and VP3 of different families of the order *Picornavirales* appear to be related, it is likely that VP1, VP2 and VP3 arose after two duplication events prior to the divergences of the *Picornavirales* families (Liljas et al. 2002). This hypothesis is further supported by our quantitative analysis, which shows that the 3D structures of VP1 and VP2, VP1 and VP3, and VP2 with VP3 of the *Dicistroviridae* capsid proteins, align with 3.14, 3.11 and 3.13 RMSD along 173, 208 and 208 residues, respectively (Table 2). The selective advantage of these duplication events might be related to a rapid assembly of the capsid or to the interactions with the cell receptors and the host immune systems.

3.4 *Enterovirus* 3C and 2A cysteine-proteases

The 3C and 2A picornains are cysteine-proteases responsible for the viral polyprotein processing. According to the MEROPS peptidase database (https://www.ebi.ac.uk/merops/ accessed July 2020) (Rawlings et al. 2018), both picornains are part of the C3 family. Based on fold similarity and catalytic triad arrangement, this family is classified alongside other serine and cysteine-protease families into clan PA. Members of this clan show the same protein fold described in chymotrypsin (family S1), which consists of two homologous six-stranded antiparallel β barrels with a catalytic triad, His-Asp-Ser (His-Asp-Cys or His-Glu-Cys in some viral proteases), located in the barrel interface (Lesk and Fordham 1996). Each barrel is made of two structural motifs composed of three antiparallel β strands connected by two loops, with strands named from A1 to F1 (N-terminal domain) and from A2 to F2 (C-terminal domain), respectively. The

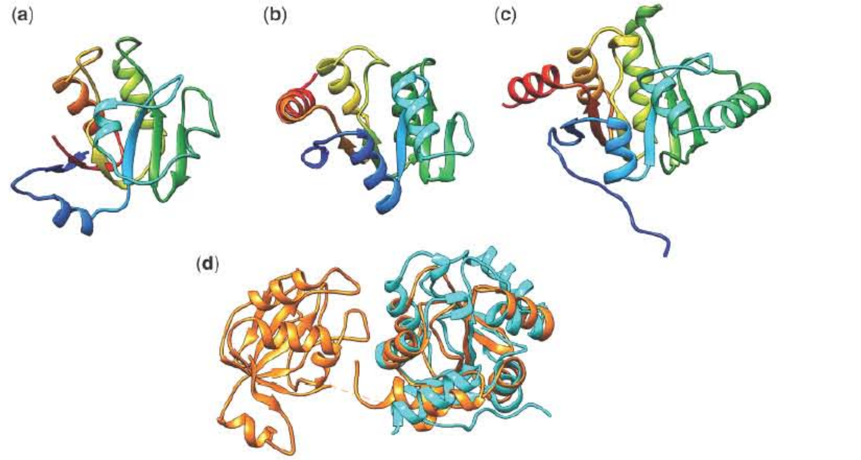


Figure 2. Tertiary structure of SUD-N (a), SUD-M (b) and X-domain (c) of SARS-CoV. Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (d) 3D alignment between SUD (orange) and X-domain (cyan) as made by FATCAT rigid. PDB IDs 2WCT and 2ACF.

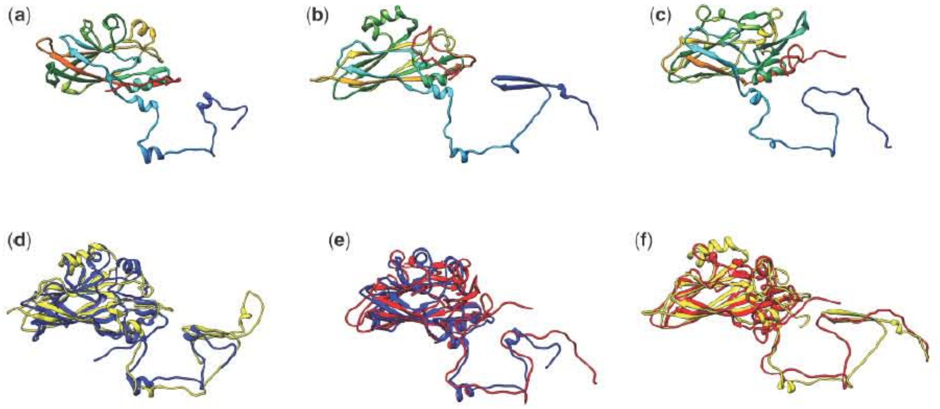


Figure 3. Tertiary structure of *Dicitroviridae* (a) VP1, (b) VP2 and (c) VP3. Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (d–f) Pairwise 3D alignments between VP1 (blue), VP2 (yellow) and VP3 (red) as made by FATCAT rigid. PDB IDs (a, b) 1B35, (c) 5CDD, (d) 1B35 and 5CDD, (e) 1B35 and 5CDD and (f) 1B35 and 3NAP.

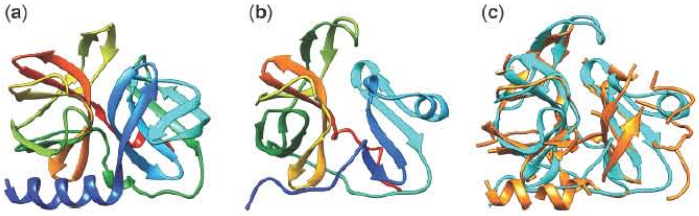


Figure 4. Tertiary structure of *Enterovirus* 3C (a) and 2A (b). Models are colored in a blue (N-terminal) to red (C-terminal) gradient. (c) 3D alignment between 3C (orange) and 2A (cyan) as made by FATCAT rigid. PDB IDs 3Q3Y and 3W95.

catalytic residues are located in different loops named accordingly with the corresponding residue. The histidine loop is located between strands C1 and D1, the aspartate loop between strands E1 and F1 and the serine loop between strands C2 and D2 (James et al. 1978; Petersen et al. 1999).

Picornains differ from chymotrypsin in that both have a shorter D1-E1 (which does not bind to calcium) and A2-B2 loops (James et al. 1978; Matthews et al. 1994). Another major difference is in the B2-C2 loop (referred as methionine loop in S1 proteases) that presents an abrupt extended turn in picornains instead of the characteristic helix found in chymotrypsin (James et al. 1978; Allaire et al. 1994; Matthews et al. 1994). The major difference between 3C and 2A is the absence of A1 and D1 strands in 2A, whose domain 1 consists of only four β stands, which makes it ∼40 resides shorter than 3C (Fig. 4). Additionally, 2A has a pair of cysteines, close to the strand A2, that mediates zinc ion coordination together with another cysteine and a histidine located in D2-E2 loop. This coordination is very similar to the one found in hepacivirin (family S29) of *Hepatitis C virus* (family *Flaviviridae*) and might play a stabilizing role such as the disulfide bond found in chymotrypsin in a similar position (Petersen et al. 1999).

It has been suggested that 3C and 2A proteases are paralogs based on a pairwise sequence alignment in which, despite the low sequence identity (13%), the catalytic residues and predicted secondary structure elements appear to be conserved (Bazan and Fletterick 1988). Based on a qualitative structure comparison, the paralogous relationship of 3C and 2A has also been suggested (Petersen et al. 1999). Our analysis adds quantitative support to the duplication hypothesis.

Information about the function and taxonomic distribution of both proteases can provide insights into the evolutionary relevance of this duplication. The 3C protease is responsible for most of the polyprotein processing, whereas 2A can only catalyze its own cleavage from the structural proteins. 3C is found in every genera of the family *Picornaviridae*, while 3C-like proteases are found in all the families of the order *Picornavirales* and in some related families such as *Caliciviridae*, *Coronaviridae* and *Potyviridae* (King et al. 2011). On the other hand, there are up to five different types of 2A proteins in the family *Picornaviridae*: 1) the 2A protease (2A^{pro}) typical of the genus *Enterovirus*; 2) the 2A^{npgp} protein typical of *Cardiovirus*, *Senecavirus*, *Aphthovirus*, *Teschovirus* and *Erbovirus*, which produces an analogous effect to the 2A^{pro} cleavage through ribosomal skipping in a conserved sequence motif NPGP; 3) the 2A$^{H\text{-}box/NC}$ protein typical of *Parechovirus*, *Kobuvirus* and *Tremovirus*, which lacks proteolytic activity and is related to a family of proteins involved in the control of cell proliferation (Hughes and Stanway 2000); 4) the

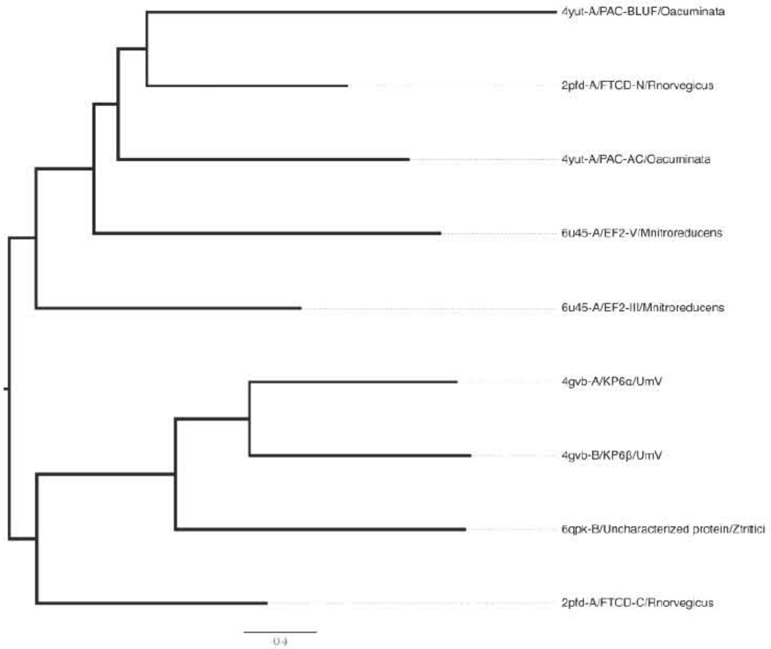

4yut-A/PAC-BLUF/Oacuminata
2pfd-A/FTCD-N/Rnorvegicus
4yut-A/PAC-AC/Oacuminata
6u45-A/EF2-V/Mnitroreducens
6u45-A/EF2-III/Mnitroreducens
4gvb-A/KP6α/UmV
4gvb-B/KP6β/UmV
6qpk-B/Uncharacterized protein/Ztritici
2pfd-A/FTCD-C/Rnorvegicus
0.4

Figure 5. Structure similarity tree supporting a close relationship between KP6α and KP6β. PDB IDs with its chain, protein names and organisms are indicated on each leaf. PAC-AC and PAC-BLUF stand for photoactivated adenylate cyclase-adenylate cyclase domain and blue light using flavin domain, respectively, FTCD equates to Formimidoyl Transferase Cyclo Deaminase amino (-N) and carboxy (-C) terminal domains and EF2 corresponds to Elongation Factor 2 domains III and V. Organisms are: UmV = *Ustilago maydis virus*, Ztritici = *Zymoseptoria tritici*, Rnorvegicus = *Rattus norvegicus*, Oacuminata = *Oscillatoria acuminata* and Mnitroreducens = *Methanoperedens nitroreducens*.

2A AIG1-like protein with possible NTPase function, located between a $2A^{npgp}$ and a $2A^{H\text{-}box/NC}$, only found in *Avihepatovirus* (Tseng et al. 2007); and 5) a 2A protein of unknown function unrelated to the previous ones only found in the genus *Hepatovirus* (King et al. 2011). This suggests that the 2A protease is a synapomorphy with a particular selective advantage on the genus *Enterovirus* (and possibly on the closely related genus *Sapelovirus*), and that the polyprotein processing activity of $2A^{pro}$, which can be replaced by the $2A^{npgp}$ or 3C, may be in fact dispensable for most picornaviruses.

Additional functions have been associated with both picornains. On the one hand, 3C has been associated to the viral replication initiation complex formation via 5′-untranslated region binding and to the host transcription inhibition through the degradation of the H3 histone, the TATA-binding protein or some transcription factors. On the other hand, $2A^{pro}$ has been associated with the host translation inhibition by means of eIF4G degradation (Bazan and Fletterick 1988; Porter 1993; Matthews et al. 1994; Petersen et al. 1999). Degradation of eIF4G allows the virus to impair the host protein translation while it takes advantage of the translation machinery through its internal ribosome entry site (IRES). It has been suggested that picornaviruses lacking a $2A^{pro}$ have a strong IRES for ribosome binding that can compete with an intact host initiation factor complex, whereas *Enterovirus* IRES binding is weak, so that these viruses depend on eIF4G inhibition to gain access to the host translation machinery (Petersen et al. 1999).

4. Discussion

Since gene homology within a genome can result from recombination and not from paralogous duplications, we performed searches against protein structure databases for each case reported here to construct structure similarity trees as a means to distinguish between the different possible scenarios. In all four cases, the trees display topologies consistent with paralogous relationships (see Figs 5 and 6 and Supplementary Figs S1 and S2). This is specially clear for the KP6 subunits KP6α and KP6β, which group together with its closest relative being an uncharacterized protein from an ascomycete fungus (Fig. 5). The capsid proteins VP1, VP2 and VP3, form three monophyletic groups, each showing a similar inner topology, which suggest two paralogous duplication events prior to the divergence of the order *Picornavirales* (Fig. 6) Although the possibility of independent gains of these proteins cannot be ruled out completely, the phylogeny depicted in Fig. 6 strongly supports their origin through gene duplication events.

It has been argued that an important difference between RNA and dsDNA viruses is the number of gene duplications (Holmes 2009). However, the newly detected cases of paralogous

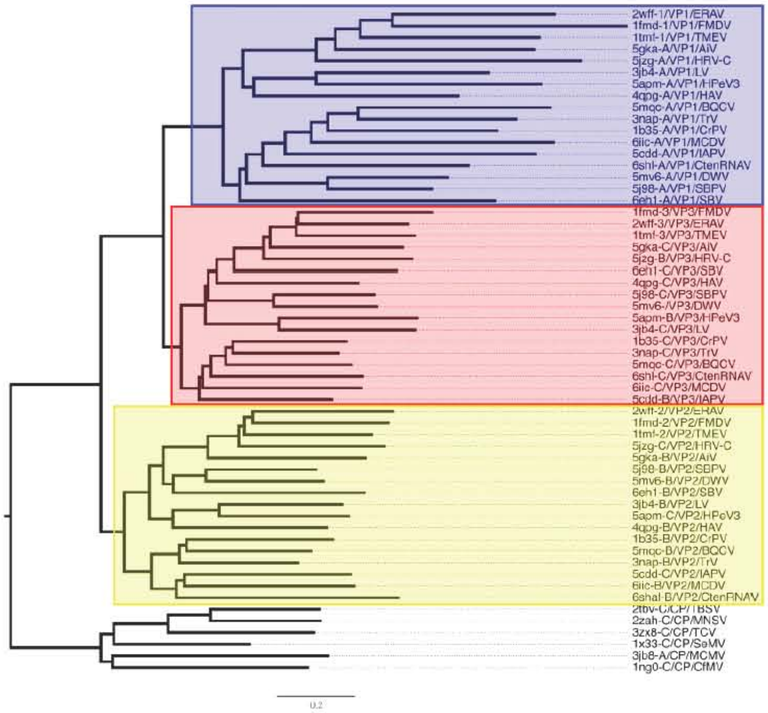


Figure 6. Structure similarity tree of proteins related to VP1, VP2 and VP3. PDB ids with its chain, protein names and organisms are indicated on each leaf. VP1, VP2 and VP3 proteins are highlighted by a blue, yellow and red box, respectively. Organisms are: ERAV = *Equine rhinitis A virus*; FMDV = *Foot and mouth disease virus*; TMEV = *Theiler's encephalomyelitis virus*; AiV = *Aichi virus*; HRV-C = *Human rhinovirus-C*; LV = *Ljungan virus*; HPeV3 = *Human parechovirus 3*; HAV = *Hepatitis A virus*; BQCV = *Black queen cell virus*; TrV = *Triatoma virus*; CrPV = *Cricket paralysis virus*; MCDV = *Mud crab dicistrovirus*; IAPV = *Israeli acute paralysis virus*; CtenRNAV = *Chaetoceros tenuissimus RNA virus*; DWV = *Deformed wing virus*; SBPV = *Slow bee paralysis virus*; SBV = *Sacbrood virus*; TBSV = *Tomato bushy stunt virus*; MNSV = *Melon necrotic spot virus*; TCV = *Turnip crinkle virus*; SeMV = *Sesbania mosaic virus*; MCMV = *Maize chlorotic mottle virus*; CfMV = *Cocksfoot mottle virus*.

proteins in RNA viruses reported here suggests that gene duplication may be a more frequent phenomenon on these viruses than previously thought. The number of detected paralogs using 3D protein comparison methodology discussed here is expected to increase as more viral protein structures are determined. Unfortunately, due to the lack of X-ray determined models, we were unable to apply our methodology to confirm the cases reported by Simon-Loriere and Holmes (2013). The addition of structural models determined with techniques other than X-ray crystallography (like NMR or Cryo-EM) may increase the size of the analyzed database.

Our inability to detect some previously reported duplications is an indication of the limits of our approach. Examples of suggested duplicated domains that are not discussed in our analysis include the shell (S) and protruding (P) domains of *Tombusviridae* capsid protein (Jones et al. 1989), as well as the P1 and P2 domains of *Hepeviridae* and *Caliciviridae* capsid proteins (Guu et al. 2009). It is worth pointing out that we also found structural similarity between the *Macrobrachium rosenbergii nodavirus* capsid P domain and the *Black beetle virus* capsid S domain (Lsas = 7.7631), both of which have a jelly-roll topology (Wery et al. 1994; Chen et al. 2019), which suggests that some nodaviruses may have undergone a duplication similar to the one suggested by Jones et al. (1989) for tombusviruses. Other similar structures that might indicate duplication events but will require further analysis are the *Porcine reproductive and respiratory syndrome virus* (*Arteriviridae*) nsp1α and nsp1β papain-like cysteine protease domains (Sun et al. 2009; Xue et al. 2010) (Lsas = 6.2778), the coronavirus 3C-like protease and nsp9 (Sutton et al. 2004) (Lsas = 13.912), the ssRNA(-) *Human respiratory syncytial virus* (*Pneumoviridae*) NS1 and matrix protein (Chatterjee et al. 2017) (Lsas = 7.4347), the retrovirus capsid N-terminal domain and C-terminal domain (Lsas = 7.0739–8.1531) and the retrovirus reverse transcriptase-ribonuclease H (RT-RNaseH) connection domain, RNaseH domain and integrase (INT) (Lsas INT-RNaseH = 5.7483–6.4967) (Malik and Eickbush 2001), although it has been suggested that the later have independent evolutionary histories (Koonin et al. 2015). Finally, It is important to note that, despite their low sequence similarity, coronavirus papain-like proteases PL1pro and PL2pro have been suggested to be paralogs (Lee et al. 1991; Herold et al. 1999; Ziebuhr et al. 2000; Ziebuhr et al. 2001). This case was not detected by our method because both sequences were clustered together by PSI-CD-HIT. Specifically, the *Swine acute diarrhea syndrome coronavirus* PL2pro (PDB: 6L5T) was selected as the representative protein for the cluster in which the *Porcine transmissible gastroenteritis coronavirus* PL1pro (PDB: 3MP2) and the *Porcine epidemic diarrhea virus* PL2pro (PDB: 6NOZ) were included with 27.404 per cent and 44.872 per cent sequence identity, respectively (Supplementary data S4).

Given the positive correlation between paralog number and genome size (Gevers et al. 2004), we would have expected to find more duplicated genes in viruses with larger genomes. For example, viruses of the family *Coronaviridae* with genomes that can reach more than 30kb (Campillo-Balderas et al. 2015) or viruses with segmented genomes which, on average, tend to be larger than non-segmented RNA genomes (Holmes 2009). However, most of the detected cases belong to monopartite viruses with genomes not larger than 20kb. This might suggest different growth mechanisms or even a sample bias. For the particular case of the coronaviruses, in which we have detected three likely paralogs, it has been suggested that their large genomes are possible because they encode a HEL domain helicase and an ExoN domain 3′-5′ exoribonuclease which are involved in RNA duplex unwinding, and proofreading and repair, respectively (Gorbalenya et al. 2006; Holmes 2009). The presence of a helicase domain could explain the number of duplicated genes detected so far in the order *Picornavirales*. Interestingly, it has been shown that large single and multi-domain protein families are less frequent in viruses compared to cellular organisms. Also, the percentage of multi-domain proteins belonging to viruses tends to be smaller than the percentage of single-domain proteins (Forslund et al 2019). Considering that the reported duplication events in RNA viruses only involve single domains of around 300 residues or less, it is possible that RNA viruses can preserve gene duplications only if the genetic redundancy is comprised of relatively small sequences. This appears to be the case presented in Willemsen et al. (2017), where the artificial insertion of the relatively small gene 2b, which codes for a redundant function, is preserved despite the predicted fitness cost of a growing genome. Although gene duplication and horizontal gene transfer imply a different homology origin, the fitness effects related to the genome size limitations are predicted to be the same. Therefore, functional redundancy may also be beneficial both after the recruitment of external sequences or following a duplication event, which is consistent with our results as well as with other gene duplication reports (Simon-Loriere and Holmes 2013) where at best we can only see slight indications of functional diversification.

Studies on RNA virus genome size increase due to recombination and/or paralogous duplications provide insights into how hypothetical cellular RNA genomes grew during early stages of cell evolution and increased their coding capacity from a small number of coding genes to the hundreds or thousands of genes that eventually led to a complex organism such as the last common ancestor (Becerra et al. 2007). Despite the obvious differences, some features of RNA virus genomes can be used as models to understand the hypothetical RNA/protein World cellular genomes. For instance, early proteinic polymerases were probably just as error prone as current viral RNA-dependent RNA polymerases (Reyes-Prieto et al. 2012; Jácome et al. 2015), whose palm subdomain is probably one of the oldest structural domains still recognizable in today's viruses and cells, and may actually be a relic from the RNA/protein World (Jácome et al. 2015).

If we add the paralogous pairs reported here to the 20 pairs listed by Simon-Loriere and Holmes (2013) at least fifty pairs distributed in twelve paralogous families composed of two to three members derived from fifteen duplication events can be considered. Although this numbers might still indicate that gene duplications are infrequent in RNA viruses, it is remarkable that gene duplications still occur and are maintained, which indicates that although genome growth tends to reduce the virus fitness (Willemsen et al. 2016), paralogous genes can be maintained whenever the acquired benefits are greater than the cost of a longer genome. Gene duplications can be conspicuous in only some RNA viruses. For instance, the three capsid proteins and the tandem repeat of VPg represent a high proportion of genes originated by gene duplication in the genome of *Foot-and-mouth disease virus*.

5. Conclusions

New cases of paralogous proteins that probably diverged early on RNA virus evolution are reported here. These cases include both subunits of the cytotoxin KP6, SARS-CoV SUD and X-domain, *Enterovirus* cysteine proteases 3C and 2A, and *Picornavirales* VP1, VP2 and VP3 viral capsids. Due to the low sequence conservation, these cases could only be confirmed by quantitative tertiary structure comparisons. The number of known paralogous proteins in RNA viruses is likely to grow as more viral protein structures are determined. Overall, our results suggest that gene duplication is a more relevant mechanism for increasing the coding capacities of RNA genomes than previously thought.

Acknowledgments

We thank the theoretical and methodological support provided by José Alberto Campillo Balderas, Ricardo Hernández Morales and Alejandro Rodrigo Jácome Ramírez. Alejandro Miguel Cisneros-Martínez is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 814975 from the Consejo Nacional de Ciencia y Tecnología (CONACYT).

Funding

This study was funded by PAPIIT-UNAM (IN214421).

Data availability

Availability of data and material: Figs S1 and S2 are available as supplementary material. Supplementary datasets (1–8) are available in a public repository: https://github.com/abb-GB/gene-duplication.git.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest

None declared.

References

Allaire, M. et al. (1994) 'Picornaviral 3C Cysteine Proteinases Have a Fold Similar to Chymotrypsin-like Serine Proteinases', *Nature*, 369: 72–6.

Allen, A., Chatt, E., and Smith, T. J. (2013) 'The Atomic Structure of the Virally Encoded Antifungal Protein, KP6', *Journal of Molecular Biology*, 425: 609–21.

Bazan, J. F., and Fletterick, R. J. (1988) 'Viral Cysteine Proteases Are Homologous to the Trypsin-like Family of Serine Proteases: Structural and Functional Implications', *Proceedings of the National Academy of Sciences of the United States of America*, 85: 7872–6.

Becerra, A. et al. (2007) 'The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains', *Annual Review of Ecology, Evolution, and Systematics* , 38: 361–79.

Berman, H. M. et al. (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28: 235–42.

Blasdell, K. R. et al. (2012) 'Kotonkan and Obodhiang Viruses: African Ephemeroviruses with Large and Complex Genomes', *Virology*, 425: 143–53.

Boyko, V. P. et al. (1992) 'Coat Protein Gene Duplication in a Filamentous RNA Virus of Plants', *Proceedings of the National Academy of Sciences of the United States of America*, 89: 9156–60.

Bulheller, B. M., and Hirst, J. D. (2009) 'DichroCalc – Circular and Linear Dichroism Online', *Bioinformatics (Oxford, England)*, 25: 539–40.

Campillo-Balderas, J. A., Lazcano, A., and Becerra, A. (2015) 'Viral Genome Size Distribution Does Not Correlate with the Antiquity of the Host Lineages', *Frontiers in Ecology and Evolution*, 3: 143.

Chandrasekar, V., and Johnson, J. E. (1998) 'The Structure of Tobacco Ringspot Virus: A Link in the Evolution of Icosahedral Capsids in the Picornavirus Superfamily', *Structure (London, England : 1993)*, 6: 157–71.

Chatterjee, S. et al. (2017) 'Structural Basis for Human Respiratory Syncytial Virus NS1-Mediated Modulation of Host Responses', *Nature Microbiology*, 2:

Chen, N. C. et al. (2019) 'The Atomic Structures of Shrimp Nodaviruses Reveal New Dimeric Spike Structures and Particle Polymorphism', *Communications Biology*, 2: 72.

Davison, A. J., Benkő, M., and Harrach, B. (2003) 'Genetic Content and Evolution of Adenoviruses', *Journal of General Virology*, 84: 2895– 2908.

Eigen, M. (1971) 'Self-Organization of Matter and the Evolution of Biological Macromolecules', *Die Naturwissenschaften* , 58: 465–523.

Fazeli, C. F., and Rezaian, M. A. (2000) 'Nucleotide Sequence and Organization of Ten Open Reading Frames in the Genome of Grapevine Leafroll-Associated Virus 1 and Identification of Three Subgenomic RNAs', *The Journal of General Virology* , 81: 605–615.

Felsenstein, J. (1989) 'PHYLIP—Phylogeny Inference Package (Version 3.2)', *Cladistics*, 5: 164–166.

Fitch, W. M., and Margoliash, E. (1967) 'Construction of Phylogenetic Trees', *Science (New York, N.Y.)*, 155: 279–284.

Forslund, S. K., Kaduk, M., and Sonnhammer, E. L. L. (2019) 'Evolution of Protein Domain Architectures', in M., Anisimova (ed.) *Evolutionary Genomics. Statistical and Computational Methods*, 2nd edn, pp 469–504. New York: Humana.

Forss, S., and Schaller, H. (1982) 'A Tandem Repeat Gene in a Picornavirus', *Nucleic Acids Research*, 10: 6441–6450.

Gago, S. et al. (2009) 'Extremely High Mutation Rate of a Hammerhead Viroid', *Science (New York, N.Y.)*, 323: 1308.

Gao, Y. et al. (2017) 'Extent and Evolution of Gene Duplication in DNA Viruses', *Virus Research*, 240: 161–165.

Gevers, D. et al. (2004) 'Gene Duplication and Biased Functional Retention of Paralogs in Bacterial Genomes', *Trends in Microbiology*, 12: 148–154.

Gorbalenya, A. E. et al. (2006) '*Nidovirales*: Evolving the Largest RNA Virus Genome', *Virus Research*, 117: 17–37.

Gubala, A. et al. (2010) 'Ngaingan Virus, a Macropod-Associated Rhabdovirus, Contains a Secondglycoprotein Gene and Seven Novel Open Reading Frames', *Virology*, 399: 98–108.

Guu, T. S. Y. et al. (2009) 'Structure of the Hepatitis E Virus-like Particle Suggests Mechanisms for Virus Assembly and Receptor Binding', *Proceedings of the National Academy of Sciences of the United States of America*, 106: 12992–12997.

Herold, J., Siddell, S. G., and Gorbalenya, A. E. (1999) 'A Human RNA Viral Cysteine Proteinase That Depends upon a Unique Zn^{2+}-Binding Finger Connecting the Two Domains of a Papain-like Fold', *The Journal of Biological Chemistry* , 274: 14918–14925.

Holm, L. (2020) 'DALI and the Persistence of Protein Shape', *Protein Science : a Publication of the Protein Society*, 29: 128–140.

Holmes, E. C. (2009) *The Evolution and Emergence of RNA Viruses*. Oxford: Oxford University Press.

—— (2011) 'What Does Virus Evolution Tell Us about Virus Origins? ', *Journal of Virology*, 85: 5247–5251.

Hughes, A. L., and Friedman, R. (2005) 'Poxvirus Genome Evolution by Gene Gain and Loss', *Molecular Phylogenetics and Evolution*, 35: 186–195.

Hughes, P. J., and Stanway, G. (2000) 'The 2A Proteins of Three Diverse Picornaviruses Are Related to Each Other and to the H-rev107 Family of Proteins Involved in the Control of Cell Proliferation', *The Journal of General Virology* , 81: 201–207.

Humphrey, W., Dalke, A., and Schulten, K. (1996) 'VMD: Visual Molecular Dynamics', *Journal of Molecular Graphics*, 14: 33–38.

Jácome, R. et al. (2015) 'Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications', *PLoS ONE*, 10: e0139001.

James, M. N. G., Delbaere, L. T. J., and Brayer, G. D. (1978) 'Amino Acid Sequence Alignment of Bacterial and Mammalian Pancreatic Serine Proteases Based on Topological Equivalences', *Canadian Journal of Biochemistry*, 56: 396–402.

Jones, E. Y., Stuart, D. I., and Walker, N. P. C. (1989) 'Structure of Tumor Necrosis Factor', *Nature*, 338: 225–228.

Kalynych, S., Pálková, L., and Plevka, P. (2016a) 'The Structure of Human Parechovirus 1 Reveals an Association of the RNA Genome with the Capsid', *Journal of Virology*, 90: 1377–1386.

—— et al. (2016b) 'Virion Structure of Iflavirus Slow Bee Paralysis Virus at 2.6-Angstrom Resolution', *Journal of Virology*, 90: 7444–7455.

Kambol, R., Kabat, P., and Tristem, M. (2003) 'Complete Nucleotide Sequence of an Endogenous Retrovirus from the Amphibian, Xenopus laevis', *Virology*, 311: 1–6.

King, A. M. Q. et al. (2011) Virus Taxonomy. *Ninth Report of the International Committee on Taxonomy of Viruses*. London: Elsevier Academic Press

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015) 'Origins and Evolution of Viruses of Eukaryotes: The Ultimate Modularity', *Virology*, 479-480: 2–25.

Kreuze, J. F., Savenkov, E. I., and Valkonen, J. P. T. (2002) 'Complete Genome Sequence and Analyses of the Subgenomic RNAs of Sweet Potato Chlorotic Stunt Virus Reveal Several

New Features for the Genus Crinivirus', *Journal of Virology*, 76: 9260–9270.

LaPierre, L. A. et al. (1999) 'Sequence and Transcriptional Analyses of the Fish Retroviruses Walleye Epidermal Hyperplasia Virus Types 1 and 2: Evidence for a Gene Duplication', *Journal of Virology*, 73: 9393–9403.

Lazcano, A. (1995) 'Cellular Evolution during the Early Archean: What Happened between the Progenote and the Cenancestor? ', *Microbiología SEM*, 11: 185–198.

Lee, H. J. et al. (1991) 'The Complete Sequence (22 Kilobases) of Murine Coronavirus Gene 1 Encoding the Putative Proteases and RNA Polymerase', *Virology*, 180: 567–582.

Lesk, A. M., and Fordham, W. D. (1996) 'Conservation and Variability in the Structures of Serine Proteinases of the Chymotrypsin Family', *Journal of Molecular Biology*, 258: 501–537.

Li, W., and Godzik, A. (2006) 'Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences', *Bioinformatics (Oxford, England)*, 22: 1658–9.

Liljas, L. et al. (2002) 'Evolutionary and Taxonomic Implications of Conserved Structural Motifs between Picornaviruses', *Archives of Virology*, 147: 59–84.

Malik, H. S., and Eickbush, T. H. (2001) 'Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses', *Genome Research*, 11: 1187–1197.

Matthews, D. A. et al. (1994) 'Structure of Human Rhinovirus 3C Protease Reveals a Trypsin-like Polypeptide Fold, RNA-Binding Site, and Means for Cleaving Precursor Polyprotein', *Cell*, 77: 761–771.

Maynard Smith, J., and Szathmáry, E. (1995) *The Major Transitions in Evolution*. Oxford: Oxford University Press.

McGeoch, D., and Davison, J. (1999) 'Molecular Evolutionary History of the Herpesviruses', in E., Domingo, R.G., Webster, H.F., Holland (eds.) *Origin and Evolution of Viruses*, pp. 441–465. London: Academic Press.

Ohno, S. (1970) *Evolution by Gene Duplication*. New York: Springer-Verlag.

Petersen, J. F. W. et al. (1999) 'The Structure of the 2A Proteinase from a Common Cold Virus: A Proteinase Responsible for the Shut-off of Host-Cell Protein Synthesis', *The EMBO Journal*, 18: 5463–5475.

Pettersen, E. F. et al. (2004) 'UCSF Chimera–a Visualization System for Exploratory Research and Analysis', *Journal of Computational Chemistry*, 25: 1605–12.

Porter, A. (1993) 'Replication and Inhibition of Host Cell Functions', *Journal of Virology*, 67: 6917–6921.

Rawlings, N. D. et al. (2018) 'The MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors in 2017 and a Comparison with Peptidases in the PANTHER Database', *Nucleic Acids Research*, 46: D624–D632.

Rambaut, A. (2014) *FigTree v1.4.2, a Graphical Viewer of Phylogenetic Trees*. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. <http://tree.bio.ed.ac.uk/software/fig tree/> (accessed February 2021).

Reanney, D. C. (1982) 'The Evolution of RNA Viruses', *Annual Review of Microbiology*, 36: 47–73.

Reyes-Prieto, F. et al. (2012) 'Coenzymes, Viruses and the RNA World', *Biochimie*, 94: 1467–1473.

Roberts, E. et al. (2006) 'MultiSeq: Unifying Sequence and Structure Data for Evolutionary Analysis', *BMC Bioinformatics*, 7: 382.

Rossmann, M. G., and Johnson, J. E. (1989) 'Icosahedral RNA Virus Structure', *Annual Review of Biochemistry*, 58: 533–73.

Russell, R. B., and Barton, G. J. (1992) 'Multiple Protein Sequence Alignment from Tertiary Structure Comparisons: Assignment of Global and Residue Confidence Levels', *Proteins* , 14: 309–323.

Sabin, C. et al. (2016) 'Structure of Aichi Virus 1 and Its Empty Particle: Clues to Kobuvirus Genome Release Mechanism', *Journal of Virology*, 90: 10800–10810.

Saikatendu, K. S. et al. (2005) 'Structural Basis of Severe Acute Respiratory Syndrome Coronavirus ADP-Ribose-1"-Phosphate Dephosphorylation by a Conserved Domain of nsP3', *Structure (London, England : 1993)*, 13: 1665–1675.

Shackelton, L. A., and Holmes, E. C. (2004) 'The Evolution of Large DNA Viruses: Combining Genomic Information of Viruses and Their Hosts', *TRENDS in Microbiology*, 12: 458–465.

Simon-Loriere, E., and —— (2013) 'Gene Duplication is Infrequent in the Recent Evolutionary History of RNA Viruses', *Molecular Biology and Evolution*, 30: 1263–1269.

Subbiah, S., Laurents, D. V., and Levitt, M. (1993) 'Structural Similarity of DNA-Binding Domains of Bacteriophage Repressors and the Globin Core', *Current Biology : Cb*, 3: 141–148.

Sun, Y. et al. (2009) 'Crystal Structure of Porcine Reproductive and Respiratory Syndrome Virus Leader Protease Nspα', *Journal of Virology*, 83: 10931–10940.

Sutton, G. et al. (2004) 'The nsp9 Replicase Protein of SARS-Coronavirus, Structure and Funcional Insights', *Structure (London, England : 1993)*, 12: 341–353.

Tan, J. et al. (2009) 'The SARS-Unique Domain (Sud) of SARS Coronavirus Contains Two Macrodomains That Bind G-Quadruplexes', *PLoS Pathogens*, 5: e1000428.

Taylor, J., and Raes, J. (2004) 'Duplication and Divergence: The Evolution of New Genes and Old Ideas', *Annual Review of Genetics*, 38: 615–643.

Tseng, C. H., Knowles, N. J., and Tsai, H. J. (2007) 'Molecular Analysis of Duck Hepatitis Virus Type 1 Indicates That It Should Be Assigned to a New Genus', *Virus Research*, 123: 190–203.

Tria, F. D. K., Landan, G., and Dagan, T. (2017) 'Phylogenetic Rooting Using Minimal Ancestor Deviation', *Nature Ecology & Evolution*, 1: 193.

Tristem, M. et al. (1990) 'Origin of Vpx in Lentiviruses', *Nature*, 347: 341–342.

Tzanetakis, I. E., and Martin, R. R. (2007) 'Strawberry Chlorotic Fleck: Identification and Characterization of a Novel Closterovirus Associated with the Disease', *Virus Research*, 124: 88–94.

——, Postman, J. D., and Martin, R. R. (2005) 'Characterization of a Novel Member of the Family Closteroviridae from Mentha Spp', *Phytopathology*, 95: 1043–1048.

Walker, P. J. et al. (1992) 'The Genome of Bovine Ephemeral Fever Rhabdovirus Contains Two Related Glycoprotein Genes', *Virology*, 191: 49–61.

Wery, J. P. et al. (1994) 'The Refined Three-Dimensional Structure of an Insect Virus at 2.8 a Resolution', *Journal of Molecular Biology*, 235: 565–586.

Willemsen, A. et al. (2016) 'Predicting the Stability of Homologous Gene Duplications in a Plant RNA Virus', *Genome Biology and Evolution*, 8: 3065–3082.

—— et al. (2017) '2b or Not 2b: Experimental Evolution of Functional Exogenous Sequences in a Plant RNA Virus', *Genome Biol. Evol*, 9: 297–310.

Xue, F. et al. (2010) 'The Crystal Structure of Porcine Reproductive and Respiratory Syndrome Virus Nonstructural Protein Nsp1β Reveals a Novel Metal-Dependent Nuclease', *Journal of Virology*, 84: 6461–6471.

Ye, Y., and Godzik, A. (2003) 'Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists', *Bioinformatics*, 19: ii246–ii255.

Zhang, J. (2003) 'Evolution by Gene Duplication: An Update', *Trends in Ecology & Evolution* , 18: 292–298.

Ziebuhr, J., Snijder, E. J., and Gorbalenya, A. E. (2000) 'Virus-Encoded Proteinases and Proteolytic Processing in the *Nidovirales*', *The Journal of General Virology* , 81: 853–879.

——, Thiel, V., and —— (2001) 'The Autocatalytic Release of a Putative RNA Virus Transcription Factor from Its Polyprotein Precursor Involves Two Paralogous Papain-like Proteases That Cleave the Same Peptide Bond', *The Journal of Biological Chemistry* , 276: 33220–33232.

Anexo 1: tablas y figuras suplementarias del artículo requisito

File S1 Full list of accession numbers for metagenomes used in this study

SRA run accesion	Sample name	Location	Description	Raw reads	Filtered reads	Publication
SRR6913561	CH2	Churince (Mexico)	Cuatro Ciénegas	7,962,496	7,186,796	Taboada et al. 2018
SRR6913558	CH4	Churince (Mexico)	Cuatro Ciénegas	9,974,898	7,440,045	Taboada et al. 2018
SRR6913571	CH5	Churince (Mexico)	Cuatro Ciénegas	8,269,314	7,218,432	Taboada et al. 2018
SRR6913570	CH9	Churince (Mexico)	Cuatro Ciénegas	4,247,137	3,584,080	Taboada et al. 2018
SRR6913569	CH10	Churince (Mexico)	Cuatro Ciénegas	7,017,080	5,534,424	Taboada et al. 2018
SRR6913568	BE	La Becerra (Mexico)	Cuatro Ciénegas	8,193,308	6,904,872	Taboada et al. 2018
SRR6913567	PR1	Pozas Rojas (Mexico)	Cuatro Ciénegas	4,805,076	4,115,469	Taboada et al. 2018
SRR6913566	PR3	Pozas Rojas (Mexico)	Cuatro Ciénegas	5,203,288	3,971,715	Taboada et al. 2018
SRR6913560	PR4	Pozas Rojas (Mexico)	Cuatro Ciénegas	13,320,976	9,924,966	Taboada et al. 2018
SRR6913559	PR7	Pozas Rojas (Mexico)	Cuatro Ciénegas	7,266,635	5,782,866	Taboada et al. 2018
SRR6913563	PR9	Pozas Rojas (Mexico)	Cuatro Ciénegas	5,908,946	5,196,793	Taboada et al. 2018
SRR001044	PA	Pozas Azules II (Mexico)	Microbialite	318,774	302,359	Desnues et al. 2008
SRR001045	RM	Rio Mesquites (Mexico)	Microbialite	394,902	328,887	Desnues et al. 2008
SRR001061	Highborne cay	Bahamas	Microbialite	162,472	147,079	Desnues et al. 2008
SRR402039	2007At1	Lake Tyrell (Australia)	Hypersaline high	1,788,587	1,133,156	Emerson et al. 2012
SRR402041	2007At2	Lake Tyrell (Australia)	Hypersaline high	4,377,964	3,628,979	Emerson et al. 2012
SRR402047	2009B	Lake Tyrell (Australia)	Hypersaline high	10,809,213	9,467,013	Emerson et al. 2012
SRR402042	2010Bt1	Lake Tyrell (Australia)	Hypersaline high	584,879	516,230	Emerson et al. 2012

SRR402043	2010Bt2	Lake Tyrell (Australia)	Hypersaline high	2,224,022	1,672,188	Emerson et al. 2012
SRR402044	2010Bt3	Lake Tyrell (Australia)	Hypersaline high	696,810	550,077	Emerson et al. 2012
SRR402045	2010Bt4	Lake Tyrell (Australia)	Hypersaline high	6,164,429	4,372,806	Emerson et al. 2012
SRR402046	2010A	Lake Tyrell (Australia)	Hypersaline high	5,907,315	4,529,547	Emerson et al. 2012
SRR001053/SRR001054	Saltern low	San Diego Bay (USA)	Hypersaline low	122,056	110,432	Dinsdale et al. 2008a
SRR001055/SRR001056	Saltern med	San Diego Bay (USA)	Hypersaline med	44,557	39,536	Dinsdale et al. 2008a
SRR001051	Tabuaeran atoll	Pacific ocean	Seawater	411,812	377,903	Dinsdale et al. 2008b
SRR001049	Palmyra atoll	Pacific ocean	Seawater	358,985	320,319	Dinsdale et al. 2008b
SRR001042	Kiritimati atoll	Pacific ocean	Seawater	329,384	282,975	Dinsdale et al. 2008b
SRR001040	Kingman atoll	Pacific ocean	Seawater	113,749	94,603	Dinsdale et al. 2008b
SRR001038	Sargasso sea	Sargasso sea	Seawater	412,745	399,625	Angly et al. 2006
SRR023774	Tampa bay	Tampa bay (USA)	Seawater	294,068	279,173	McDaniel et al. 2008
SRR001047	Tilapia pond 1105	Kent SeaTech (USA)	Freshwater	316,182	266,567	Dinsdale et al. 2008a
SRR001075	Tilapia pond 0506	Kent SeaTech (USA)	Freshwater	62,245	59,508	Dinsdale et al. 2008a
SRR001076	Prebead pond 0506	Kent SeaTech (USA)	Freshwater	69,785	66,600	Dinsdale et al. 2008a
ERR019477	Lake Pavin	Lake Pavin (France)	Freshwater	684,228	556,373	Roux et al. 2012
ERR019478	Lake Bourget	Lake Bourget (France)	Freshwater	597,693	557,531	Roux et al. 2012

SPECIES-LEVEL TAXONOMIC ASSIGNMENT	EARLY-LATE	EARLY-DEPTH	LATE-DEPTH
MICROVIRIDAE SP.	0.0015	0.0004	0.2566
CIRCOVIRIDAE SP.	0.0081	0.0106	0.2842
MICROVIRUS SP.	0.0153	0.0157	0.3708
PROKARYOTIC DSDNA VIRUS SP.	0.0231	0.0064	0.1933
UNCULTURED MARINE PHAGE	0.0394	0.0339	0.4209
HALOVIRUS HGTV-1	0.0199	0.0017	0.3799
HALORUBRUM PHAGE GNF2	0.0224	0.0004	0.3006
HALORUBRUM VIRUS HRTV-28	0.0327	0.0001	0.1236
ENVIRONMENTAL HALOPHAGE EHP-32	0.0292	0.0013	0.0932
ENVIRONMENTAL HALOPHAGE EHP-2	0.0353	0.0005	0.2949
ENVIRONMENTAL HALOPHAGE EHP-15	0.0372	0.0008	0.4660
ENVIRONMENTAL HALOPHAGE EHP-31	0.0375	0.0065	0.1457
ARCHAEAL BJ1 VIRUS	0.0390	0.0106	0.2793
ENVIRONMENTAL HALOPHAGE EHP-34	0.0396	0.0002	0.2539
ENVIRONMENTAL HALOPHAGE EHP-9	0.0413	0.0014	0.1511
ENVIRONMENTAL HALOPHAGE EHP-28	0.0443	0.0001	0.3291
HALORUBRUM PHAGE CGPHI46	0.0445	0.0212	0.3167
ENVIRONMENTAL HALOPHAGE EHP-6	0.0448	0.0001	0.1952
HALOVIRUS HRTV-4	0.0475	0.0015	0.1773
ENVIRONMENTAL HALOPHAGE EHP-14	0.0496	0.0194	0.2808
SYNECHOCOCCUS PHAGE S-SCSM1	0.0039	0.0363	0.0445
ENVIRONMENTAL HALOPHAGE EHP-11	0.0283	0.0035	0.0427
ENVIRONMENTAL HALOPHAGE EHP-20	0.0329	0.0061	0.0301
HALORUBRUM VIRUS HRTV-29	0.0375	0.0001	0.0443
HALOVIRUS HCTV-2	0.0411	0.0004	0.0017
HALOVIRUS HHTV-2	0.0534	0.0011	0.0145
ENVIRONMENTAL HALOPHAGE EHP-12	0.0569	0.0002	0.0297
HALOVIRUS HHTV-1	0.0570	0.0065	0.0163
ENVIRONMENTAL HALOPHAGE EHP-16	0.0718	0.0116	0.0241
ENVIRONMENTAL HALOPHAGE EHP-24	0.0908	0.0025	0.0030
PODOVIRIDAE SP. CTPVR23	0.1054	0.0127	0.0177
ENVIRONMENTAL HALOPHAGE EHP-38	0.2002	0.0001	0.0002

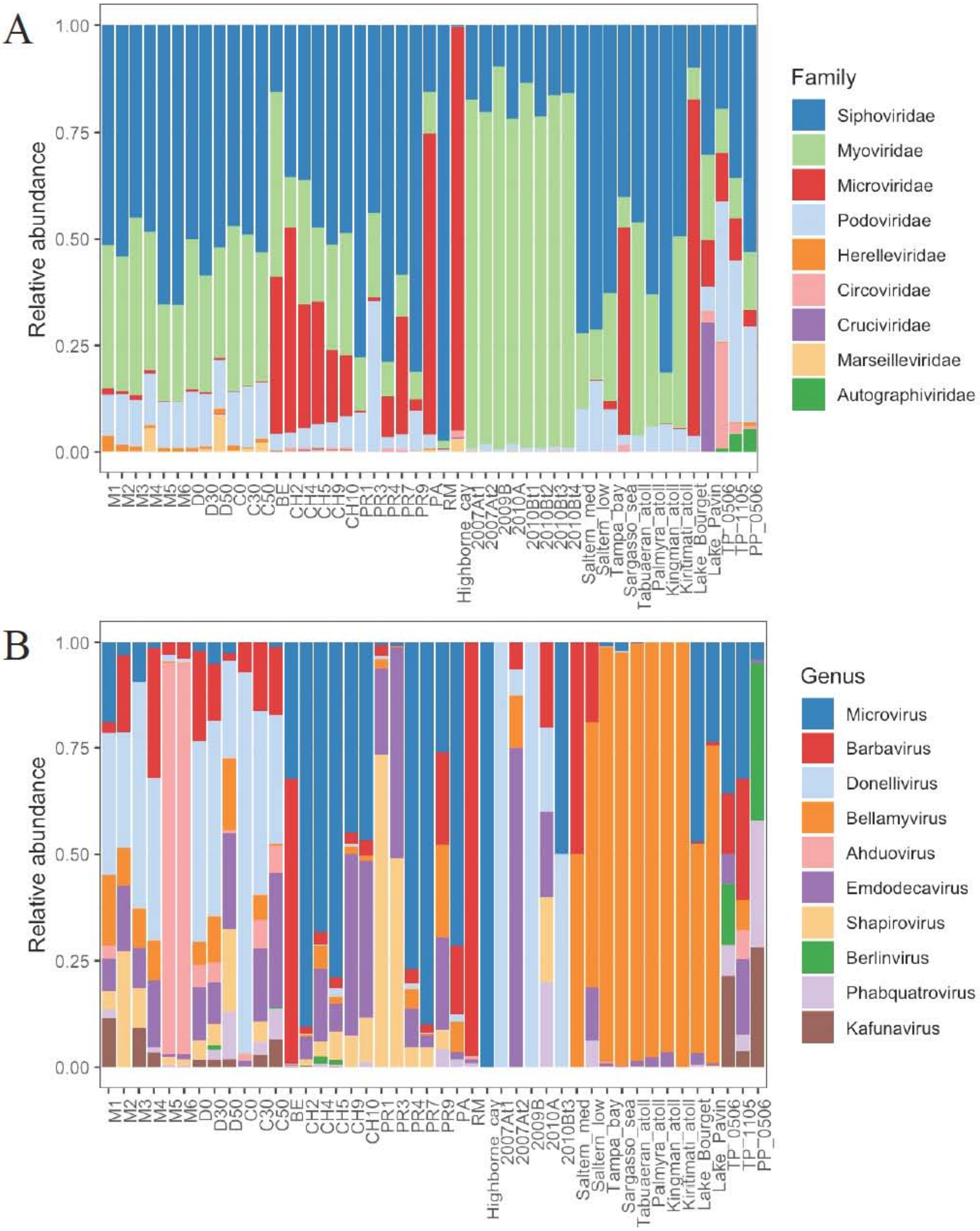
Supplemental table 1 Species-level t-Student p-values between early (n = 3), late (n = 3) and depth samples (n = 3). Only significant p-values are colored. Blue and red text designate a significatively greater or lower relative abundance for the reference sample, respectively.

SPECIES-LEVEL TAXONOMIC ASSIGNMENT	SURFACE	DEPTHS	P
SIPHOVIRIDAE SP.	25.8329	17.4080	2.0055E-05
UNCULTURED CAUDOVIRALES PHAGE	22.7780	17.4134	0.00984391
MYOVIRIDAE SP.	18.8575	11.2481	3.7883E-05
PROKARYOTIC DSDNA VIRUS SP.	7.2790	5.4325	0.00535573
BACTERIOPHAGE SP.	4.7722	3.7516	0.04271066
HALOVIRUS HSTV-1	0.3212	5.5070	0.00026996
HALOVIRUS HHTV-1	0.0552	4.9855	0.00377766
HALOVIRUS HGTV-1	0.5087	3.7021	0.00016193
UNCULTURED MARINE VIRUS	1.3306	0.9097	0.04508303
CRASS-LIKE VIRUS SP.	1.2643	0.6321	0.00290447
HALORUBRUM PHAGE GNF2	0.1797	1.7171	2.8321E-05
HALORUBRUM VIRUS HRTV-28	0.1510	1.6146	7.647E-07
UNCULTURED MARINE PHAGE	1.5131	0.2267	0.0249969
HALOVIRUS HCTV-2	0.0500	1.7151	0.00118913
ENVIRONMENTAL HALOPHAGE EHP-28	0.1048	1.3580	9.9806E-06
HALORUBRUM PHAGE CGPHI46	0.0654	1.2633	0.00123251
ARCHAEAL BJ1 VIRUS	0.0880	1.1582	0.00044929
HALOVIRUS HHTV-2	0.0340	1.0875	0.00017474
HALORUBRUM VIRUS HRTV-29	0.0836	0.7230	1.6945E-06
PONTIMONAS PHAGE PHIPSAL1	0.0669	0.5484	0.00308692
HALOVIRUS HRTV-4	0.0697	0.5177	2.1183E-05
ENVIRONMENTAL HALOPHAGE EHP-14	0.0320	0.4955	0.00103097
HALOFERAX TAILED VIRUS 1	0.0396	0.4519	1.8172E-05
ENVIRONMENTAL HALOPHAGE EHP-20	0.0405	0.4418	0.0017202
ENVIRONMENTAL HALOPHAGE EHP-34	0.0378	0.4302	5.4202E-06
ENVIRONMENTAL HALOPHAGE EHP-31	0.0239	0.3461	0.00024614
MICROVIRIDAE SP.	0.2977	0.0489	0.00123602
ENVIRONMENTAL HALOPHAGE EHP-15	0.0263	0.2294	0.00013617
ENVIRONMENTAL HALOPHAGE EHP-9	0.0222	0.2270	2.1266E-05
ENVIRONMENTAL HALOPHAGE EHP-11	0.0196	0.2131	0.00034844
MICROVIRUS SP.	0.1641	0.0299	0.01894391
ENVIRONMENTAL HALOPHAGE EHP-32	0.0204	0.1645	3.5596E-05
ENVIRONMENTAL HALOPHAGE EHP-12	0.0060	0.1192	3.4201E-06
ENVIRONMENTAL HALOPHAGE EHP-36	0.0050	0.0793	0.00038805
ENVIRONMENTAL HALOPHAGE EHP-2	0.0043	0.0570	1.8936E-05
ENVIRONMENTAL HALOPHAGE EHP-30	0.0197	0.0411	0.01863032
PROCHLOROCOCCUS PHAGE P-TIM68	0.0344	0.0048	0.00637372
ENVIRONMENTAL HALOPHAGE EHP-6	0.0010	0.0251	1.1884E-06
SYNECHOCOCCUS VIRUS BELLAMY	0.0189	0.0057	0.0377621
ENVIRONMENTAL HALOPHAGE EHP-16	0.0027	0.0205	0.00726656
BORDETELLA VIRUS PHB04	0.0035	0.0174	0.0462241
ENVIRONMENTAL HALOPHAGE EHP-24	0.0005	0.0194	0.01393725
PODOVIRIDAE SP. CTPVR23	0.0019	0.0182	0.03736158
ENVIRONMENTAL HALOPHAGE EHP-38	0.0009	0.0121	0.00169676
SPHAEROTILUS PHAGE VB_SNAP-R1	0.0005	0.0039	0.04272228

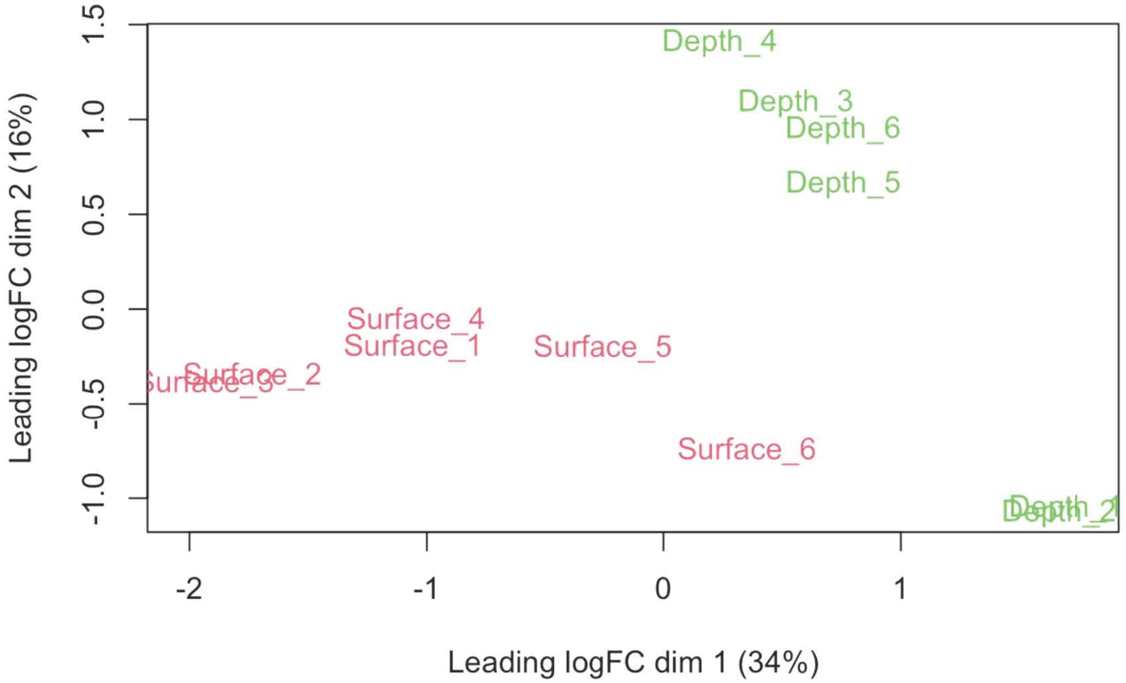
Supplemental table 2 Species-level average relative abundance and t-Student p-values of OTUs significatively more abundant (blue text) or less abundant (red text) in surface samples (n = 6) compared to deep samples *sensu lato* (n = 6).

Sample	Run	Location	Sample Type	Reads	Reference
Great_salt_lake	SRR10846467	Bridger bay, Great salt lake, Utah	Hypersaline microbial mat	19,922,338	Kanik et al. 2020
Hot_lake	SRR5271190	Hot Lake, Washington, USA	Hypersaline microbial mat	30,318,005	Lindemann et al. 2013
Tristomo_elos12	ERR6290777	Tristomo bay (Karpathos, Greece)	Hypersaline microbial mat	111,720,365	Pavloudi et al. 2022
Tristomo_elos1	ERR6290772	Tristomo bay (Karpathos, Greece)	Hypersaline microbial mat	134,529,087	Pavloudi et al. 2022
Tristomo_elos7	ERR6290775	Tristomo bay (Karpathos, Greece)	Hypersaline microbial mat	144,236,134	Pavloudi et al. 2022
DK32S	SRR9330145	Habor Lake (Inner Mongolia Autonomous Region, China)	Soda lake	67,353,198	Zhao et al. 2020
HC22W	SRR9330148	Hutong Qagan Lake (Inner Mongolia Autonomous Region, China)	Soda lake	47,829,334	Zhao et al. 2020
HC26S	SRR9330142	Hutong Qagan Lake (Inner Mongolia Autonomous Region, China)	Soda lake	56,302,065	Zhao et al. 2020
HC5W	SRR9330150	Hutong Qagan Lake (Inner Mongolia Autonomous Region, China)	Soda lake	52,132,576	Zhao et al. 2020
Wadi_El-Natrun	ERR1770058	Wadi El-Natrun, Egypt	Soda lake	19,989,051	ZeinEldin et al. 2023

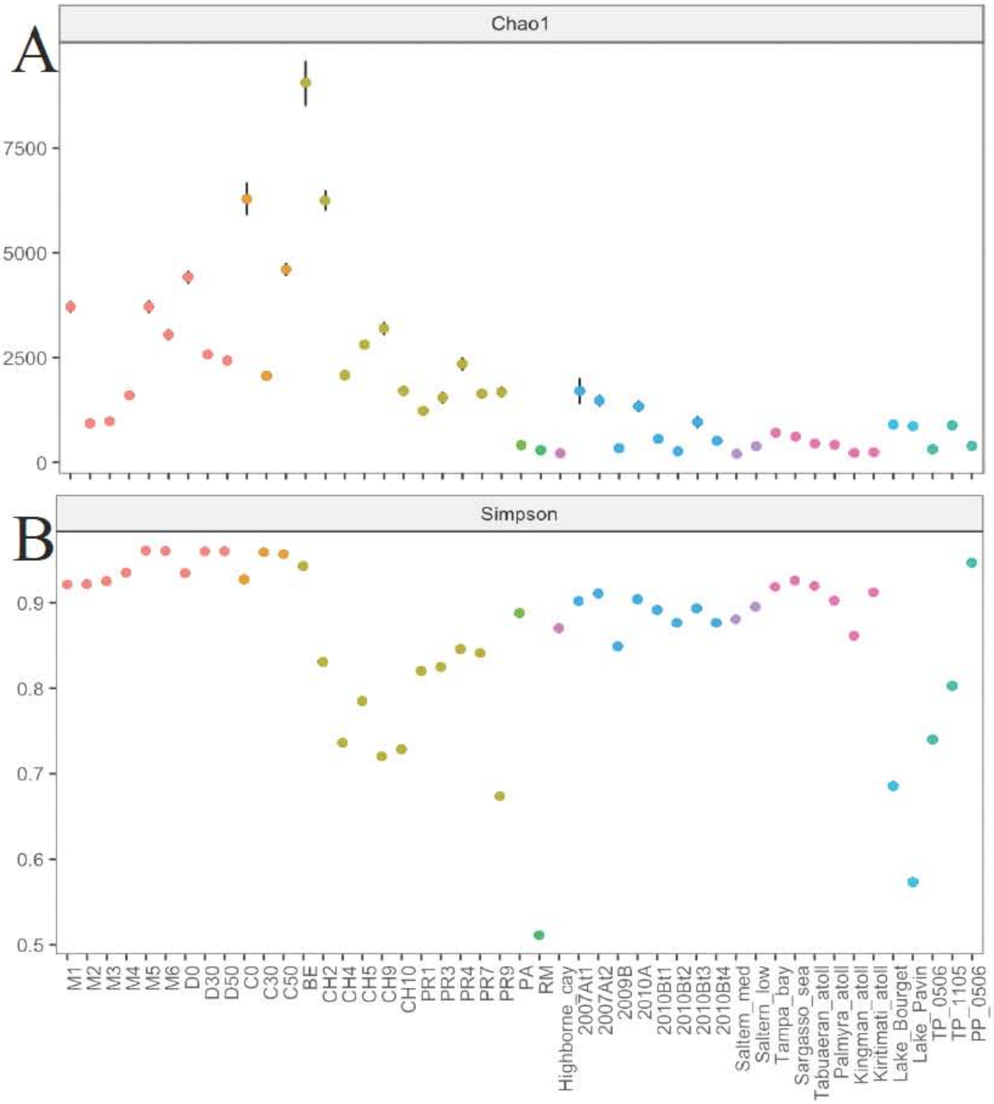
Supplemental table 3 Samples used to compare AD viral community with that of hypersaline microbial mats and soda lakes.



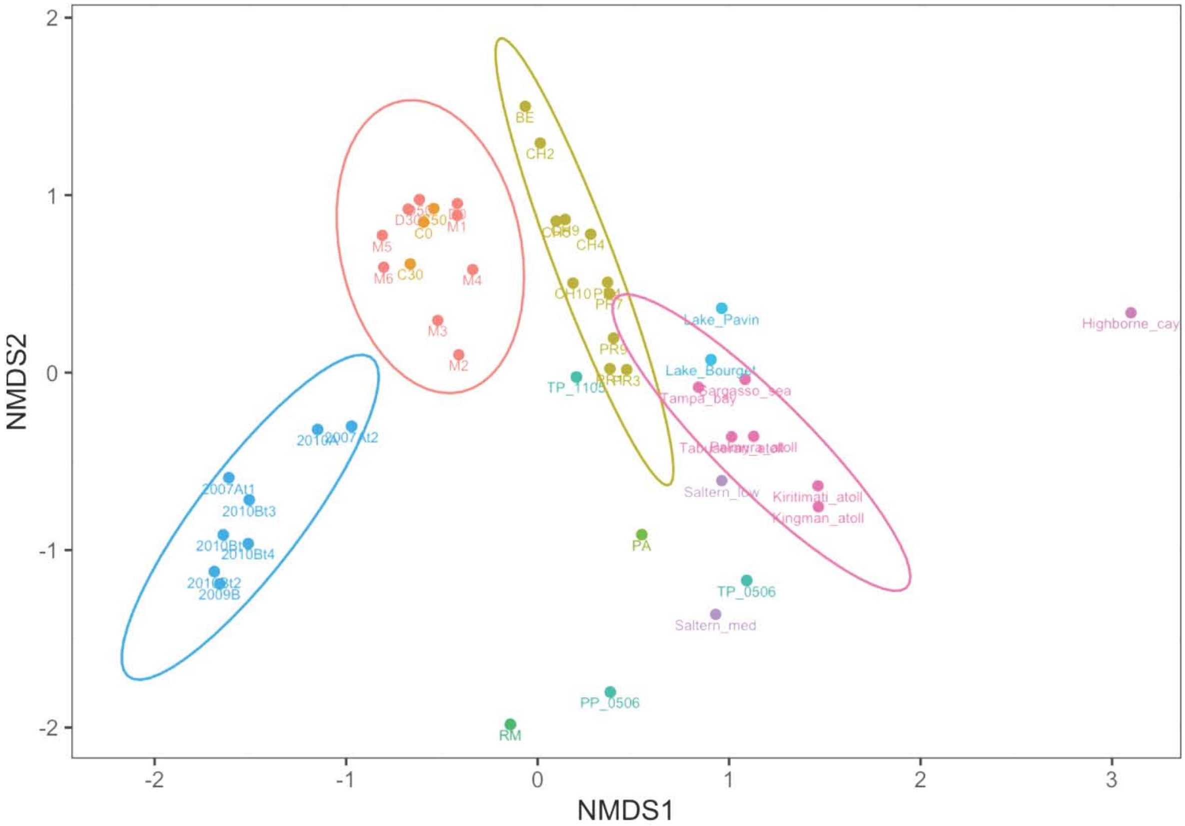
Supplementary figure 2 Relative abundance of reads assigned to viral families (A) and genera (B) for 12 AD metagenomes and 35 metagenomes from other environments (see Supplementary table 1).



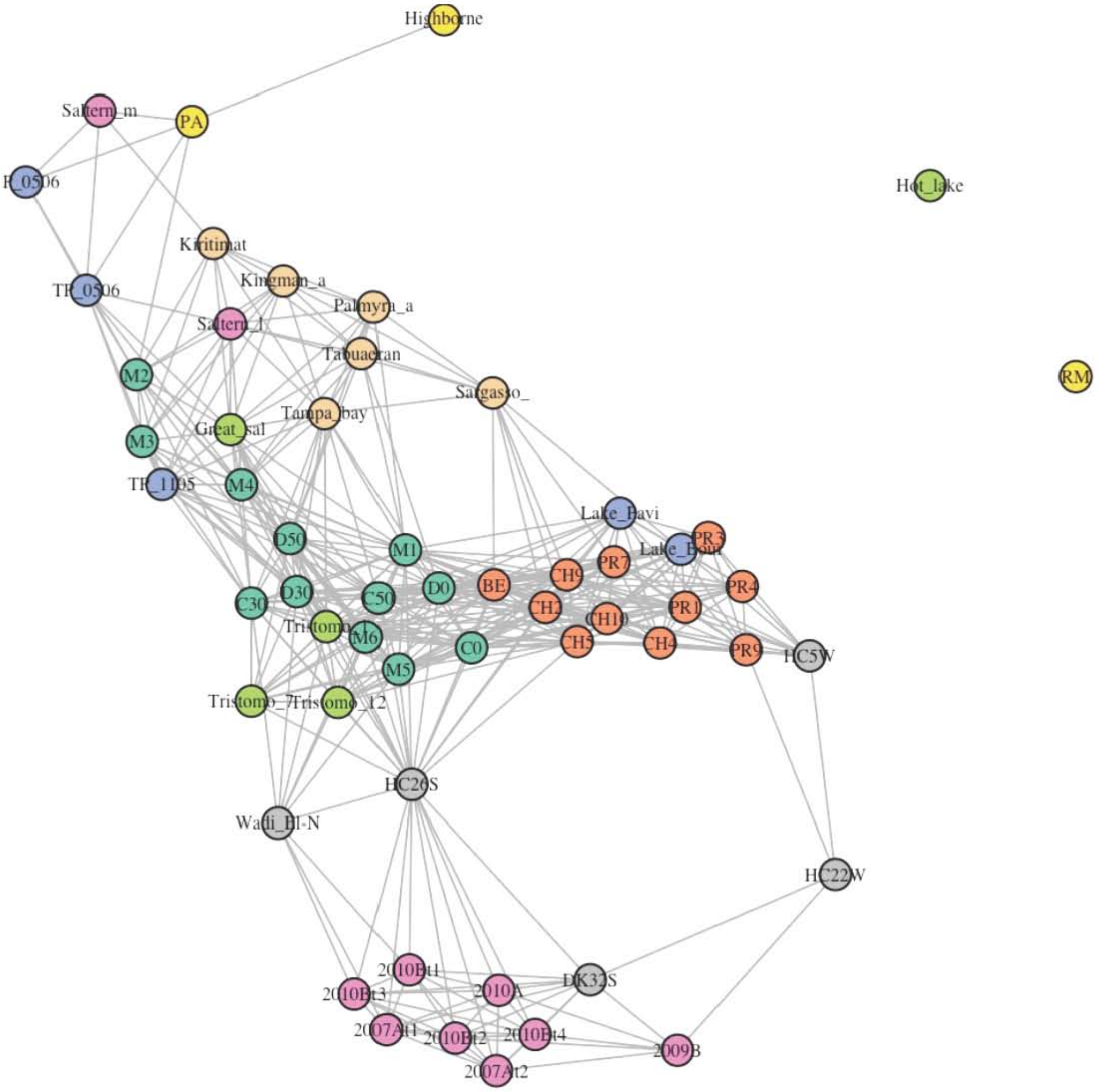
Supplementary figure 3 MDS analysis of all OTUs normalized counts. Surface_1 = M1, Surface_2 = M2, Surface_3 = M3, Surface_4 = M4, Surface_5 = D0, Surface_6 = C0, Depth_1 = M5, Depth_2 = M6, Depth_3 = D30, Depth_4 = D50, Depth_5 = C30, Depth_6 = C50.



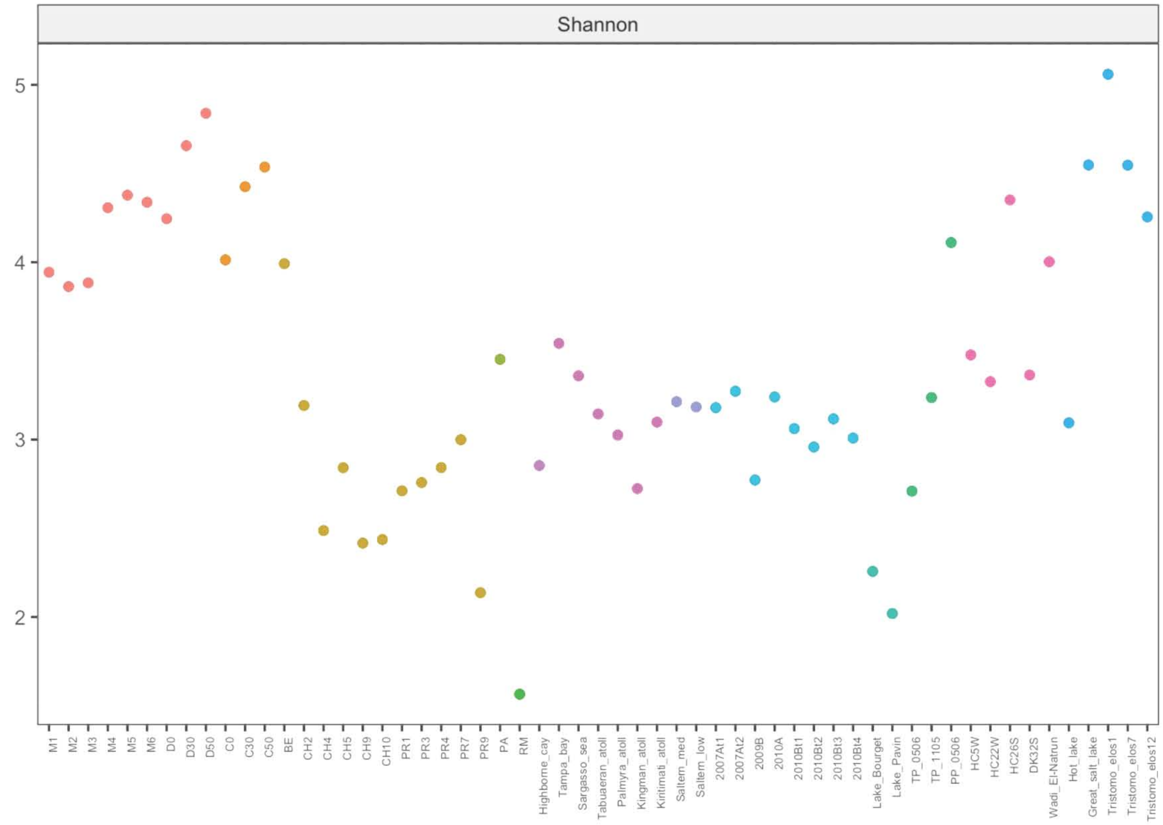
Supplemental figure 4 Chao1 (A) and Simpson (B) diversity indexes for 12 AD metagenomes and 35 metagenomes from other environments (see Supplemental File 1). AD viromes are represented by red points. Other CCB viromes (PR and CH) are represented by olive green points. Sea viromes are represented by pink points. High hypersaline viromes are represented by blue points.



Supplemental Figure 5 NMDS analysis of Bray-Curtis dissimilarities among 47 viromes. ellipses represent 95% confidence interval for a multivariate t distribution. AD viromes are represented by red points agglomerated inside a red ellipse. Other CCB viromes (PR and CH) are represented by olive green points aggregated inside an olive green ellipse. Sea viromes are represented by pink points inside a pink ellipse. High hypersaline viromes are represented by blue points inside a blue ellipse.



Supplemental Figure 7 OTU level Bray-Curtis similarity network showing 25% (above 3rd quartile) of the strongest similarities. AD viromes are represented by jade green circles. Other hypersaline microbial mats are represented by lime green circles. Soda lakes viromes are represented by grey circles. Other CCB viromes (PR and CH) are represented by orange circles. Ocean viromes are represented by beige circles. High hypersaline viromes are represented by pink circles within.



Supplemental Figure 8 OTU level alpha diversity index (Shannon) including five viromes from hypersaline microbial mats (blue points) and soda lakes (magenta points), respectively.

Anexo 2: tabla suplementaria del artículo 2

S4_File. Full list of host predictions for metagenomic assembled viruses from Archaean Domes.

Metagenome	Contig name	Short name	Size	Completeness	Coverage	CRISPR	CrisprCustomDB	CrisprOpenDB	PHP	RaFAH	Host biology
C50	NODE_1_length_42463_cov_31.019534_cutoff_20_type_linears	C50N1L42	42463	Full-length	102.20%	**d__Bacteria;p__Desulfobacterota;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfohalobiaceae;g__Desulfovermiculus;s__**	NA	NA	**d__Bacteria;p__Desulfobacterota;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfohalobiaceae;g__**	**Bacteria; Thermodesulfobacteriota; Desulfovibrionia; Desulfovibrionales; Desulfovibrionaceae; Desulfovibrio**	Sulfate reduction; haloalkaline
M5	NODE_2_length_43874_cov_79.277025_cutoff_54_type_circular	M5N2L438	43874	Full-length	100.70%	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	NA	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	Halophilic
M6	NODE_1_length_43908_cov_85.795619_cutoff_54_type_circular	M6N1L439	43908	Full-length	100.70%	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	NA	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Nitrococcales;f__Halorhodospiraceae;g__Halorhodospira**	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	Halophilic
C30	NODE_1_length_64430_cov_14.133757_cutoff_10_type_circular	C30N1L64	64430	Full-length	101.40%	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Thiohalorhabdales/Thiohalospirales;f__Thiohalorhabdaceae/Thiohalospiraceae;g__Thiohalorhabdus/Thiohalospira**	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Thiohalorhabdales;f__Thiohalorhabdaceae;g__Thiohalorhabdus; Criterion 3: Multiple hosts matching same number of regions. Host with spacer**	NA	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Thiohalorhabdales;f__Thiohalorhabdaceae;g__Thiohalorhabdus**	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	Hypersaline lake

							closest to the 5' end					
C0	NODE_5_length_50610_cov_20.948735_cutoff_5_type_linears	C0N5L506	50610	Full-length	117.20%	**d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfobacterales;f__SURF-3;g__**	NA	NA	**d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfobacterales;f__SURF-3;g__**	Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiaceae; Clostridium	Sulfate reduction
M1	NODE_5_length_60786_cov_6.133055_cutoff_5_type_linears	M1N5L607	60786	Full-length	140.70%	**d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfobacterales;f__SURF-3;g__;s__**	**d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfobacterales;f__SURF-3;g__;s__**	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Alteromonadales; Pseudoalteromonadaceae; Pseudoalteromonas	Sulfate reduction
C0	NODE_1_length_39461_cov_45.003458_cutoff_42_type_circular	C0N1L394	39461	Full-length	94.90%	**d__Bacteria;p__Desulfobacterota;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfohalobiaceae;g__Desulfovermiculus;s__**	NA	NA	**d__Bacteria;p__Desulfobacterota;c__Desulfovibrionia;o__Desulfovibrionales;f__Desulfohalobiaceae;g__**	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides	Sulfate reduction, haloalkaline
M4	NODE_1_length_64231_cov_24.449473_cutoff_10_type_circular	M4N1L642	64231	Full-length	100.20%	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;f__Chromatiaceae;g__Halochromatium;s__**	**d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;f__Chromatiaceae;g__Halochromatium;s__**	d_Bacteria; p_Pseudomonadota; c_Betaproteobacteria; o_Nitrosomonadales ;f_Thiobacillaceae;g_Thiobacillus	NA	Bacteria; Pseudomonadota; Betaproteobacteria; Neisseriales; Neisseriaceae; Kingella	Hypersaline microbial mat
C0	NODE_2_length_45847_cov_20.668504_cutoff_5_type_circular	C0N2L458	45847	Full-length	108.30%	NA	NA	**d_Bacteria;c_Gammaproteobacteria;o_Oceanospirillales;f_Halomonadacea,g_Halomonas**	**d__Bacteria;c__Gammaproteobacteria;o__Chromatiales;f__Chromatiaceae;g__Halochromatium**	Bacteria; Pseudomonadota; Betaproteobacteria; Rhodocyclales; Zoogloeaceae; Thauera	Hypersaline microbial mat
M5	NODE_6_length_41552_cov_26.743005_cutoff_10_type_linears	M5N6L415	41552	Full-length	135.90%	NA	NA	NA	**d__Archaea;c__Hadarchaeia;o__;f__;g__**	**Archaea; Euryarchaeota; Halobacteria; Halobacteriales; Haloarculaceae; Haloarcula**	Archaea
M6	NODE_2_length_524	M6N2L524	52409	Full-length	96.50%	NA	NA	NA	**d__Archaea;c__Halob**	**Archaea; Euryarcha**	Archaea

	09_cov_4.634750_cutoff_5_type_circular								**acteria;o__Halobacteriales;f__Haloferacaceae;g__Halorubrum**	**eota; Halobacteria; Halobacteriales; Haloarculaceae; Haloarcula**	
M1	NODE_1_length_79019_cov_26.416443_cutoff_15_type_circular	M1N1L790	79019	Full-length	160.20%	NA	NA	**d_Bacteria;p_Bacillota;c_Halanaerobiia,o_Halanaerobiales;f_Halanaerobiaceae;g_Halanaerobium**	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiaceae; Clostridium	Haloalkaliphilic
D30	NODE_111_length_14048_cov_5.062280_cutoff_5_type_linears	D30N111L	14048	Full-length	97.60%	NA	NA	NA	**d__Archaea;c__Archaeoglobi;o__Archaeoglobales;f__Archaeoglobaceae;g__JdFR-22**	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	Archaea
D30	NODE_2_length_48831_cov_6.604550_cutoff_0_type_circular	D30N2L48	48831	Full-length	111.50%	NA	NA	NA	**d__Archaea;c__Bathyarchaeia;o__B26-1;f__SOJC01;g__**	Bacteria; Bacillota; Negativicutes; Veillonellales; Veillonellaceae; Veillonella	Archaea
D30	NODE_115_length_13220_cov_7.609257_cutoff_5_type_linears	D30N115L	13220	Full-length	91.80%	NA	NA	NA	**d__Archaea;c__Nanoarchaeia;o__UBA583;f__;g__**	Bacteria; Bacillota; Bacilli; Bacillales; Bacillaceae; Bacillus	Archaea
M4	NODE_1_length_42454_cov_7.179956_cutoff_0_type_circular	M4N1L424	42454	Full-length	115.80%	NA	NA	NA	**d__Bacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhodomicrobiaceae;g__Dichotomicrobium**	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Tannerellaceae; Parabacteroides	Thermohalophilic
D30	NODE_1_length_56803_cov_15.492801_cutoff_10_type_circular	D30N1L56	56803	Full-length	169.50%	NA	NA	NA	**d__Bacteria;c__Aminicenantia; o__Aminicenantales; f__Aminicenantaceae;g__**	Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiaceae; Clostridium	Hydrothermal vents; groundwater
M3	NODE_8_length_36483_cov_6.129690_cu	M3N8L364	36483	Full-length	96.90%	NA	NA	NA	**d__Bacteria;c__Anaerolineae;o__SBR1031**	Bacteria; Pseudomonadota; Gammaproteobacteri	Marine sediments

	toff_5_type_linears								**;f__A4b;g__**	a; Vibrionales ; Vibrionaceae; Vibrio	
M1	NODE_5_length_60833_cov_5.310381_cutoff_5_type_circular	M1N5L608	60833	Full-length	94.90%	NA	NA	NA	**d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038**	Bacteria; Fusobacteriota; Fusobacteriia; Fusobacteriales; Fusobacteriaceae; Fusobacterium	Marine sediments
M5	NODE_3_length_64571_cov_11.528738_cutoff_5_type_circular	M5N3L645	64571	Full-length	100.70%	NA	NA	NA	**d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038**	Bacteria; Pseudomonadota; Betaproteobacteria; Neisseriales; Neisseriaceae; Kingella	Marine sediments
C50	NODE_2_length_80914_cov_16.413866_cutoff_0_type_circular	C50N2L80	80914	Full-length	121.00%	NA	NA	NA	**d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038**	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales ; Vibrionaceae; Vibrio	Marine sediments
D50	NODE_2_length_80914_cov_21.102999_cutoff_0_type_circular	D50N2L80	80914	Full-length	121.00%	NA	NA	NA	**d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038**	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales ; Vibrionaceae; Vibrio	Marine sediments
M5	NODE_8_length_40405_cov_28.342445_cutoff_5_type_circular	M5N8L404	40405	Full-length	104.00%	NA	NA	NA	**d__Bacteria;c__Bipolaricaulia;o__;f__;g__**	Bacteria; Fusobacteriota; Fusobacteriia; Fusobacteriales; Leptotrichiaceae; Leptotrichia	Geothermal brine
M6	NODE_4_length_40405_cov_34.307091_cutoff_10_type_circular	M6N4L404	40405	Full-length	104.00%	NA	NA	NA	**d__Bacteria;c__Bipolaricaulia;o__;f__;g__**	Bacteria; Fusobacteriota; Fusobacteriia; Fusobacteriales; Leptotrichiaceae; Leptotrichia	Geothermal brine
D30	NODE_50_length_360	D30N50L3	36058	Full-length	95.90%	NA	NA	NA	**d__Bacteria;c__Bipol**	Bacteria; Pseudomo	Geothermal brine

	58_cov_5.844508_cutoff_5_type_linears								**aricaulia;o__;f__;g__**	nadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	
C50	NODE_1_length_90473_cov_7.769973_cutoff_5_type_circular	C50N1L90	90473	Full-length	100.60%	NA	NA	NA	**d__Bacteria;c__Chitinivibrionia;o__Chitinivibrionales;f__;g__**	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	Haloalkaliphilic
M1	NODE_25_length_4627_cov_32.936889_cutoff_0_type_circular	M1N25L46	4627	Full-length	102.20%	NA	NA	NA	**d__Bacteria;c__Chitinivibrionia;o__Chitinivibrionales;f__;g__**	Bacteria; Chlamydiota; Chlamydiia; Chlamydiales; Chlamydiaceae; Chlamydia	Haloalkaliphilic
D30	NODE_6_length_39700_cov_6.053066_cutoff_0_type_circular	D30N6L39	39700	Full-length	125.00%	NA	NA	NA	**d__Bacteria;c__Chitinivibrionia;o__Chitinivibrionales;f__;g__**	Bacteria; Pseudomonadota; Alphaproteobacteria; Sphingomonadales; Erythrobacteraceae; Porphyrobacter	Haloalkaliphilic
M1	NODE_22_length_26794_cov_10.205760_cutoff_5_type_linears	M1N22L26	26794	Full-length	166.80%	NA	NA	NA	**d__Bacteria;c__Chitinivibrionia;o__Chitinivibrionales;f__;g__**	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	Haloalkaliphilic
M5	NODE_4_length_59279_cov_10.527083_cutoff_5_type_circular	M5N4L592	59279	Full-length	153.20%	NA	NA	NA	**d__Bacteria;c__Gammaproteobacteria;o__Halothiobacillales;f__Halothiobacillaceae;g__**	Bacteria; Bacillota; Clostridia; Eubacteriales; Oscillospiraceae; Faecalibacterium	Halotolerant
M6	NODE_2_length_59279_cov_11.768596_cutoff_0_type_circular	M6N2L592	59279	Full-length	153.20%	NA	NA	NA	**d__Bacteria;c__Gammaproteobacteria;o__Halothiobacillales;f__Halothiobacillaceae;g__**	Bacteria; Bacillota; Clostridia; Eubacteriales; Oscillospiraceae; Faecalibacterium	Halotolerant

M5	NODE_7_length_41695_cov_11.639699_cutoff_5_type_circular	M5N7L416	41695	Full-length	106.90%	NA	NA	NA	**d__Bacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Wenzhouxiangellaceae;g__Wenzhouxiangella**	Bacteria; Pseudomonadota; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia	Marine sediments
M6	NODE_3_length_41722_cov_12.145931_cutoff_10_type_circular	M6N3L417	41722	Full-length	107.00%	NA	NA	NA	**d__Bacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Wenzhouxiangellaceae;g__Wenzhouxiangella**	Bacteria; Pseudomonadota; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia	Marine sediments
M1	NODE_1_length_52138_cov_19.099479_cutoff_10_type_circular	M1N1L521	52138	Full-length	109.60%	NA	NA	NA	**d__Bacteria;c__Gammaproteobacteria;o__XJ16;f__Halofilaceae;g__Halofilum**	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	Marine solar saltern
M5	NODE_28_length_50934_cov_6.678273_cutoff_5_type_linears	M5N28L50	50934	Full-length	93.80%	NA	NA	NA	d__Bacteria;c__Gemmatimonadetes;o__KS3-K002;f__;g__	**Archaea; Euryarchaeota; Halobacteria; Halobacteriales; Haloarculaceae; Haloarcula**	Archaea
M6	NODE_4_length_51115_cov_7.045756_cutoff_0_type_circular	M6N4L511	51115	Full-length	94.20%	NA	NA	NA	d__Bacteria;c__Gemmatimonadetes;o__KS3-K002;f__;g__	**Archaea; Euryarchaeota; Halobacteria; Halobacteriales; Haloarculaceae; Haloarcula**	Archaea
C0	NODE_2_length_195648_cov_5.103687_cutoff_5_type_linears	C0N2L195	195648	Full-length	94.60%	NA	NA	NA	**d__Bacteria;c__Halanaerobiia;o__Halanaerobiales;f__;g__**	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	Halophilic
M4	NODE_3_length_52731_cov_6.678294_cutoff_5_type_linears	M4N3L527	52731	Full-length	123.60%	NA	NA	NA	**d__Bacteria;c__Phycisphaerae; o__Phycisphaerales; f__SM1A02;g__**	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales;	Marine

										Pseudomonadaceae; Pseudomonas	
M1	NODE_3_length_46169_cov_19.868099_cutoff_15_type_circular	M1N3L461	46169	Full-length	118.10%	NA	NA	NA	**d__Bacteria;c__Rhodothermia;o__Rhodothermales;f__;g__**	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	Halophilic, thermophilic
D30	NODE_26_length_53399_cov_8.198209_cutoff_5_type_linears	D30N26L5	53399	Full-length	117.30%	NA	NA	NA	**d__Bacteria;c__Thermotogae;o__Petrotogales;f__;g__**	Bacteria; Fusobacteriota; Fusobacteriia; Fusobacteriales; Fusobacteriaceae; Fusobacterium	Thermophilic
C0	NODE_1_length_56747_cov_3.923755_cutoff_0_type_circular	C0N1L567	56747	Full-length	92.40%	NA	NA	NA	**d__Bacteria;c__Halanaerobiia;o__Halanaerobiales;f__CSSED10-376;g__**	Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiaceae; Clostridium	Halophilic
C30	NODE_1_length_45057_cov_49.892566_cutoff_26_type_circular	C30N1L45	45057	Full-length	95.30%	NA	NA	NA	NA	**Bacteria; Actinomycetota; Actinomycetes; Micromonosporales; Micromonosporaceae; Salinispora**	Marine sediments
M1	NODE_6_length_53575_cov_7.391034_cutoff_5_type_linears	M1N6L535	53575	Full-length	124.00%	NA	NA	NA	NA	**Bacteria; Bacillota; Clostridia; Eubacteriales; Desulfotomaculaceae; Desulfotomaculum**	Sulfate reduction
M1	NODE_8_length_48312_cov_8.535063_cutoff_5_type_circular	M1N8L483	48312	Full-length	110.30%	NA	NA	NA	NA	**Bacteria; Deinococcota; Deinococci; Thermales; Thermaceae; Thermus**	Thermophilic
M5	NODE_19_length_71312_cov_5.456206_cutoff_5_type_linears	M5N19L71	71312	Full-length	129.00%	NA	NA	NA	NA	**Bacteria; Pseudomonadota; Alphaproteobacteria;**	Oligotrophic

										Caulobacterales; Caulobacteraceae; Caulobacter	
M6	NODE_6_length_71444_cov_6.303911_cutoff_5_type_linears	M6N6L714	71444	Full-length	183.70%	NA	NA	NA	NA	**Bacteria; Pseudomonadota; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Caulobacter**	Oligotrophic
D50	NODE_3_length_71897_cov_4.543235_cutoff_0_type_circular	D50N3L71	71897	Full-length	124.40%	NA	NA	NA	d__Bacteria;c__Brocadiae;o__SM23-32;f__SM23-32;g__B1Sed10-231	Bacteria; Pseudomonadota; Alphaproteobacteria; Hyphomicrobiales; Methylobacteriaceae; Methylobacterium	-
M1	NODE_2_length_68187_cov_48.986012_cutoff_15_type_circular	M1N2L681	68187	Full-length	94.00%	NA	NA	d_Bacteria;p_Bacteroidota;c_Bacteroidia;o_Bacteroidales;f_Prolixibacteraceae;g_Prolixibacter	NA	Bacteria; Pseudomonadota; Alphaproteobacteria; Hyphomicrobiales; Nitrobacteraceae; Bradyrhizobium	-
M4	NODE_1_length_42773_cov_8.077569_cutoff_5_type_circular	M4N1L427	42773	Full-length	102.90%	NA	NA	d_Bacteria;p_Proteobacteria;c_Gammaproteobacteria;o_Enterobacteriales;f_Enterobacteriaceae;g_Klebsiella	d__Bacteria;c__SZUA-567;o__SZUA-567;f__;g__	Bacteria; Actinomycetota; Actinomycetes; Pseudonocardiales; Pseudonocardiaceae; Actinoalloteichus	-
C50	NODE_68_length_13115_cov_18.546119_cutoff_5_type_linears	C50N68L1	13115	Full-length	91.10%	NA	NA	NA	d__Bacteria;c__Bacteroidia;o__Bacteroidales;f__UBA12077;g__	Bacteria; Bacillota; Bacilli; Bacillales; Bacillaceae; Bacillus	-
C50	NODE_1_length_47015_cov_33.747611_cutoff_26_type_circular	C50N1L47	47015	Full-length	108.40%	NA	NA	NA	d__Bacteria;c__Dehalococcoidia;o__GIF9;f__AB-539-J10;g__	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-

D30	NODE_34_length_46980_cov_5.620152_cutoff_5_type_linears	D30N34L4	46980	Full-length	93.10%	NA	NA	NA	d__Bacteria;c__Dehalococcoidia;o__SZUA-161;f__SpSt-899;g__	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
D50	NODE_1_length_47665_cov_6.446843_cutoff_5_type_circular	D50N1L47	47665	Full-length	109.90%	NA	NA	NA	d__Bacteria;c__Dehalococcoidia;o__SZUA-161;f__SpSt-899;g__	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
C0	NODE_11_length_5637_cov_3.112704_cutoff_0_type_circular	C0N11L56	5637	Full-length	188.80%	NA	NA	NA	d__Bacteria;c__JS1;o__SB-45;f__UBA6794;g__	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
D50	NODE_5_length_51040_cov_26.375189_cutoff_0_type_circular	D50N5L51	51040	Full-length	117.70%	NA	NA	NA	d__Bacteria;c__Myxococcia;o__SLRQ01;f__SLRQ01;g__SKWY01	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	-
M1	NODE_1_length_51634_cov_39.537092_cutoff_33_type_circular	M1N1L516	51634	Full-length	116.60%	NA	NA	NA	d__Bacteria;c__SLGR01;o__;f__;g__	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	-
M1	NODE_12_length_37362_cov_9.851215_cutoff_5_type_circular	M1N12L37	37362	Full-length	91.40%	NA	NA	NA	d__Bacteria;c__UBA6919;o__UBA6919;f__;g__	Bacteria; Bacillota; Clostridia; Eubacteriales; Clostridiaceae; Clostridium	-
M1	NODE_10_length_42603_cov_12.018457_cutoff_5_type_circular	M1N10L42	42603	Full-length	114.90%	NA	NA	NA	NA	Bacteria; Actinomycetota; Actinomycetes; Actinomycetales; Actinomycetaceae;	-

										Actinomyces	
M1	NODE_3_length_88546_cov_20.232235_cutoff_5_type_circular	M1N3L885	88546	Full-length	156.70%	NA	NA	NA	NA	Bacteria; Bacillota; Bacilli; Bacillales; Bacillaceae; Bacillus	-
C0	NODE_2_length_37444_cov_40.296005_cutoff_26_type_circular	C0N2L374	37444	Full-length	107.70%	NA	NA	NA	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
D0	NODE_1_length_65573_cov_65.222428_cutoff_20_type_circular	D0N1L65a	65573	Full-length	102.30%	NA	NA	NA	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
D0	NODE_1_length_65573_cov_65.222428_cutoff_26_type_circular	D0N1L65b	65573	Full-length	102.30%	NA	NA	NA	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
M5	NODE_2_length_65995_cov_6.684445_cutoff_5_type_circular	M5N2L659	65995	Full-length	102.90%	NA	NA	NA	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
M6	NODE_1_length_65426_cov_7.789093_cutoff_5_type_circular	M6N1L654	65426	Full-length	102.00%	NA	NA	NA	NA	Bacteria; Bacillota; Clostridia; Eubacteriales; Lachnospiraceae; Roseburia	-
M1	NODE_13_length_34857_cov_13.795854_cutoff_5_type_circular	M1N13L34	34857	Full-length	102.50%	NA	NA	NA	NA	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	-
M5	NODE_29_length_49271_cov_8.653182_cutoff_5_type_linears	M5N29L49	49271	Full-length	118.50%	NA	NA	NA	NA	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	-
M6	NODE_11_length_485	M6N11L48	48589	Full-length	116.90%	NA	NA	NA	NA	Bacteria; Bacteroido	-

	89_cov_10.254137_cutoff_5_type_linears									ta; Bacteroidia; Bacteroidales; Prevotellaceae; Prevotella	
M1	NODE_14_length_32962_cov_11.571707_cutoff_5_type_circular	M1N14L32	32962	Full-length	94.80%	NA	NA	NA	NA	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Rikenellaceae; Alistipes	-
D50	NODE_1_length_160425_cov_7.452501_cutoff_5_type_linears	D50N1L16	160425	Full-length	111.40%	NA	NA	NA	NA	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Tannerellaceae; Parabacteroides	-
M1	NODE_5_length_72713_cov_5.595955_cutoff_0_type_circular	M1N5L727	72713	Full-length	182.30%	NA	NA	NA	NA	Bacteria; Cyanobacteriota; Cyanophyceae; Pleurocapsales; Dermocarpellaceae; Stanieria	-
D50	NODE_4_length_65433_cov_14.144244_cutoff_0_type_circular	D50N4L65	65433	Full-length	100.40%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia	-
C50	NODE_6_length_67694_cov_5.028313_cutoff_5_type_linears	C50N6L67	67694	Full-length	105.60%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax	-
M1	NODE_1_length_37453_cov_28.998366_cutoff_26_type_circular	M1N1L374	37453	Full-length	108.00%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Moraxellales; Moraxellaceae; Acinetobacter	-

C50	NODE_2_length_48505_cov_12.809996_cutoff_5_type_circular	C50N2L48	48505	Full-length	111.90%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
M1	NODE_53_length_3023_cov_2.668508_cutoff_0_type_circular	M1N53L30	3023	Full-length	113.90%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
M1	NODE_6_length_58440_cov_11.943820_cutoff_5_type_circular	M1N6L584	58440	Full-length	97.10%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
M6	NODE_1_length_49572_cov_31.868986_cutoff_10_type_circular	M6N1L495	49572	Full-length	98.10%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas	-
C50	NODE_1_length_44652_cov_28.530803_cutoff_20_type_circular	C50N1L44	44652	Full-length	112.20%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	-
C50	NODE_2_length_44626_cov_28.532057_cutoff_26_type_circular	C50N2L44	44626	Full-length	112.20%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	-
D0	NODE_3_length_383	D0N3L383	38384	Full-length	94.80%	NA	NA	NA	NA	Bacteria; Pseudomo	-

	84_cov_4.891549_cutoff_0_type_circular									nadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	
M2	NODE_1_length_47843_cov_19.902108_cutoff_15_type_circular	M2N1L478	47843	Full-length	114.60%	NA	NA	NA	NA	Bacteria; Pseudomonadota; Gammaproteobacteria; Vibrionales; Vibrionaceae; Vibrio	-
C0	NODE_12_length_5513_cov_100.202934_cutoff_0_type_circular	C0N12L55	5513	Full-length	100.50%	NA	NA	NA	d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	-
D0	NODE_11_length_5513_cov_395.868734_cutoff_0_type_circular	D0N11L55	5513	Full-length	98.90%	NA	NA	NA	d__Bacteria;c__Bacteroidia;o__Bacteroidales;f__BBW3;g__UBA5261	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	-
M1	NODE_21_length_5513_cov_502.249722_cutoff_0_type_circular	M1N21L55	5513	Full-length	100.50%	NA	NA	NA	d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	-
M2	NODE_8_length_5513_cov_6647.304307_cutoff_0_type_circular	M2N8L551	5513	Full-length	100.50%	NA	NA	NA	d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	-
M5	NODE_13_length_5513_cov_405.167843_cutoff_0_type_circular	M5N13L55	5513	Full-length	100.50%	NA	NA	NA	d__Bacteria;c__Anaerolineae;o__SBR1031;f__A4b;g__J038	Bacteria; Pseudomonadota; Gammaproteobacteria;	-

										Enterobacterales; Enterobacteriaceae; Escherichia	
M6	NODE_18_length_5513_cov_1483.672670_cutoff_0_type_circular	M6N18L55	5513	Full-length	98.90%	NA	NA	NA	d__Bacteria;c__Bacteroidia;o__Bacteroidales;f__BBW3;g__UBA5261	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia	-

Anexo 3: figuras suplementarias del artículo anexo

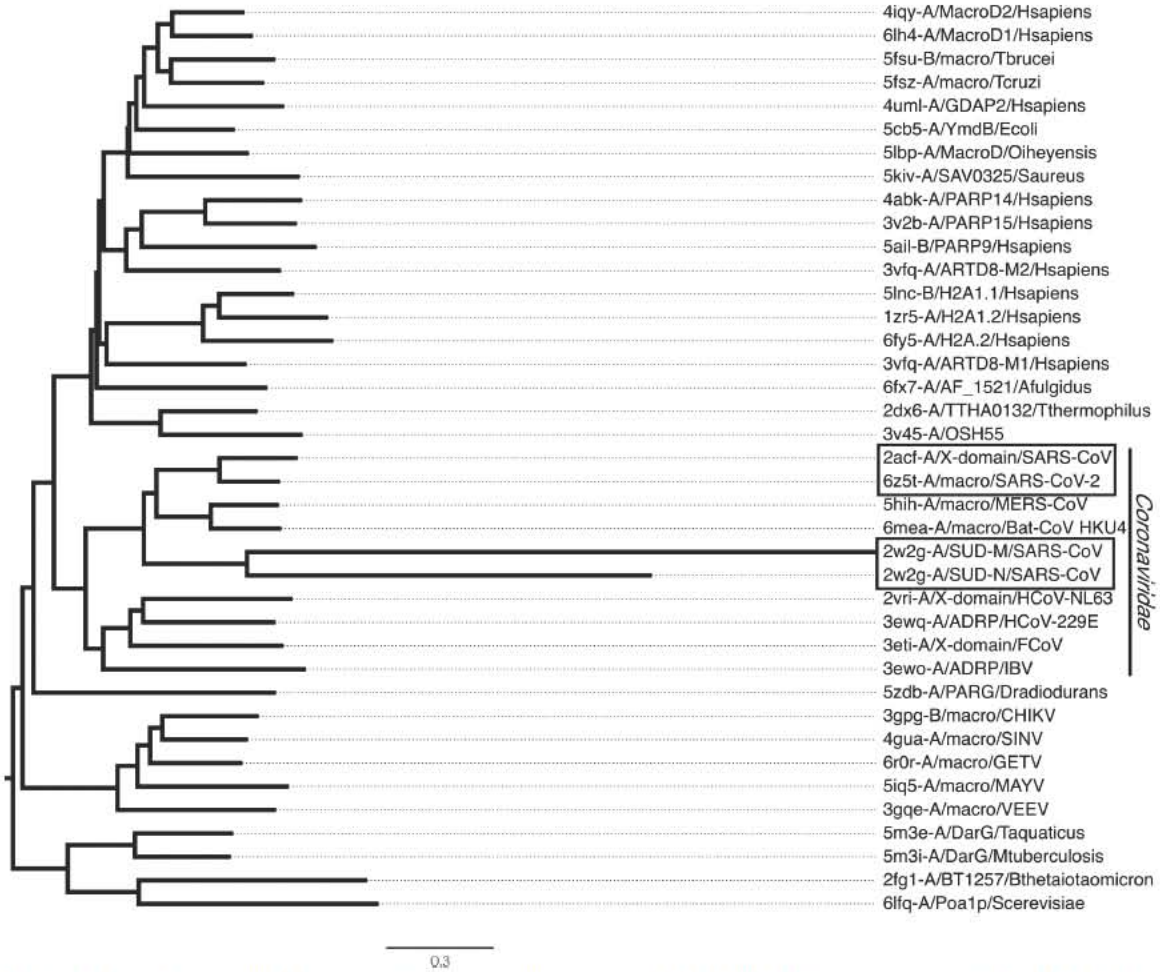


Figure S1. Structure similarity tree of protein related to SUD-N, SUD-M and X-domain. PDB ids with its chain, protein names and organisms are indicated on each leaf. Protein abbreviations correspond to: MacroD2=O-acetyl-ADP-ribose deacetylase; MacroD1=ADP-ribose glycohydrolase; macro=macrodomain; GDAP2=ganglioside-induced differentiation-associated protein 2; YmdB=O-acetyl-ADP-ribose deacetylase; MacroD=MacroD-type macrodomain; SAV0325=Protein-ADP-ribose hydrolase; PARP14=Poly [ADP-ribose] polymerase 14; PARP15=Poly [ADP-ribose] polymerase 15; PARP9=Poly [ADP-ribose] polymerase 9; ARTD8=ADP-ribosyl-transferase diphtheria toxin-like 8; H2A=histone H2A; AF_1521=ADP-ribosylglutamate; TTHA0132=hypothetical protein; OSH55=Serine hydrolase; ADRP=ADP-ribose-1”-monophosphatase; PARG=Poly ADP-ribose glycohydrolase; DarG=Appr-1-p processing domain; BT1257=conserved hypothetical protein; Poa1p=ADP-ribose 1''-phosphate phosphatase. Organisms are: Hsapiens=*Homo sapiens*; Tbrucei=*Tripanosoma brucei;* Tcruzi=*Tripanosoma cruzi*; Ecoli=*Escherichia coli*; Oiheyensis=*Oceanobacillus iheyensis*; Saureus=*Staphylococcus aureus*; Afulgidus=*Archaeoglobus fulgidus*; Tthermophilus=*Thermus thermophilus*; Dradiodurans=*Deinococcus radiodurans*; Taquaticus=*Thermus aquaticus*; Mtuberculosis=*Mycobacterium tuberculosis*; Bthetaiotamicron=*Bacteroides thetaiotamicron*; Scerevisiae=*Saccharomyces cerevisiae*; SARS-CoV=*Severe acute respiratory syndrome-related coronavirus*; MERS-CoV=*Middle East respiratory syndrome-related coronavirus*; Bat-CoV HKU4=*Bat coronavirus HKU4*; HCoV-NL63=*Human coronavirus NL63*; HCoV-229E=*Human coronavirus 229E*; FCoV=*Feline coronavirus*; IBV=*Infectious bronchitis virus*; CHIKV=*Chikungunya virus*; SINV=*Sindbis virus*; GETV=*Getah virus*; MAYV=*Mayaro virus*; VEEV=*Venezuelan equine encephalitis virus*.

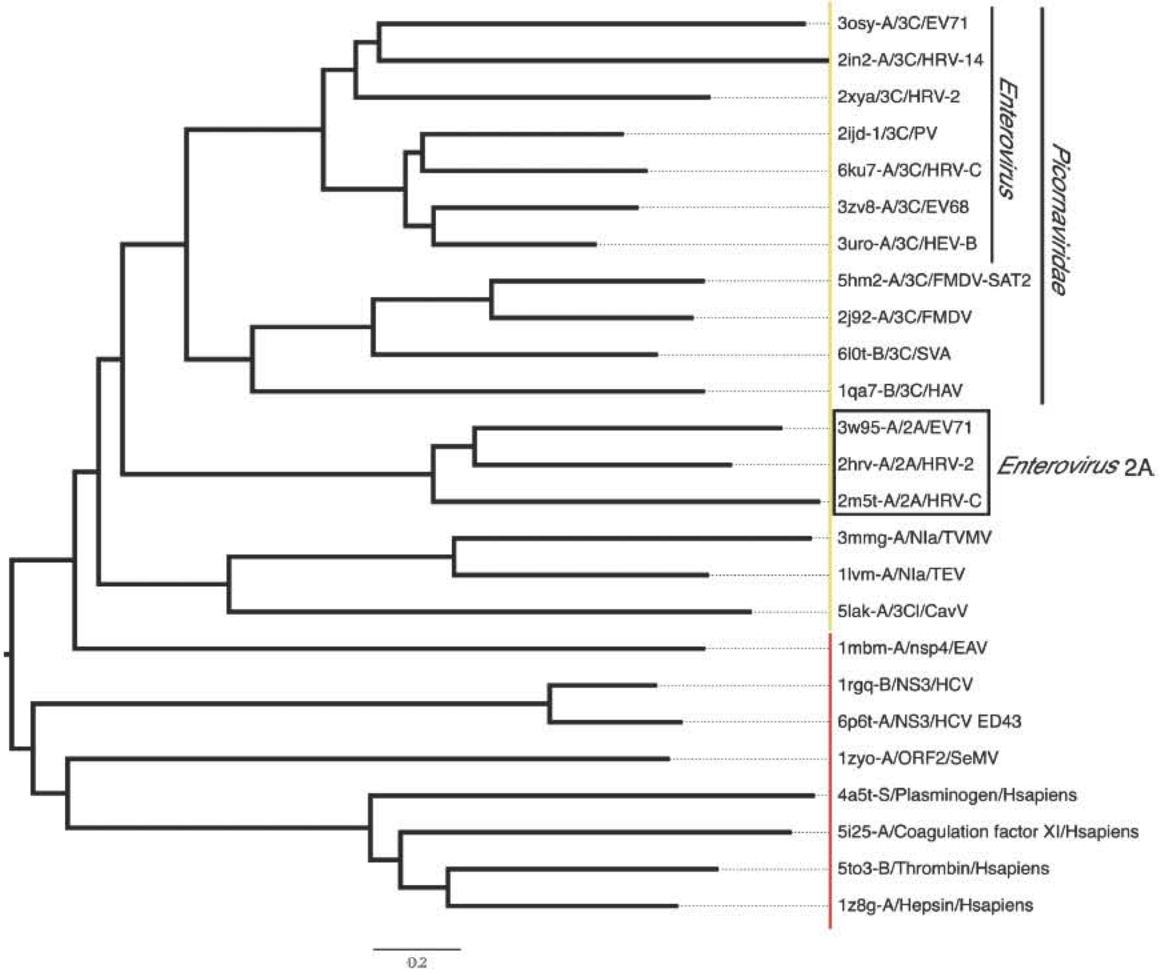


Figure S2. Structure similarity tree of proteins related to 3C and 2A proteases. PDB ids with its chain, protein names and organisms are indicated on each leaf. The yellow bar denotes cysteine proteases whereas the red line indicates serine proteases. Organisms are: Hsapiens=*Homo sapiens*; EV71=*Enterovirus 71*; HRV-14=*Human rhinovirus 14*; HRV-2=*Human rhinovirus 2*; PV=*Polio virus*; HRV-C=*Human rhinovirus C*; EV68=*Enterovirus 68*; HEV-B=*Human enterovirus B*; FMDV=*Foot and mouth disease virus*; SVA=*Senecavirus A*; HAV=*Hepatitis A virus*; TVMV=*Tobacco vein mottling virus*; TEV=*Tobacco etch virus*; CavV=*Cavally virus*; EAV=*Equine arteritis virus*; HCV=*Hepatitis C virus*; SeMV=*Sesbania mosaic virus*.

Referencias

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves predictions of hosts from metagenomically-derived viral sequences. *Nucleic Acids Research, 45*(1), 39-53.

Alcaraz, L. D., Olmedo, G., Bonilla, G., Cerritos, R., Hernández, G., Cruz, A., Ramírez, E., Putonti, C., Jiménez, B., Martínez, E., López, V., Arvisu, J. L., Ayala, F., Razo, F., Caballero, J., Siefert, J., Eguiarte, L., Vielle, J-P., Martínez, O., … Herrera-Estrella, L. (2008). The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *PNAS, 105*(15), 5803-5808.

Alcocer, J., & Hammer, U. T. (1998). Saline lake ecosystems of Mexico. *Aquatic Ecosystem Health and Management, 1*, 291-315.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol., 215*(3), 403-410.

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data* (Version 0.11.9). Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angly, F. E., Felts, B., Breibart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., Hatnes, M., Kelley, S., Liu, H., Mahaffy, J. M., Mueller, J. E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C. A., & Rohwer, F. (2006). The Marine Viromes of Four Oceanic Regions. *PLoS Biol., 4*, e368.

Antipov, D., Raiko, M., Lapidus, A., & Pevzner, P. A. (2020). METAVIRALSPADES: assembly of viruses from metagenomic data. *Bioinformatics, 36*(14), 4126-4129.

Banciu, H. L., & Sorokin, D. Y. (2013). Adaptation in Haloalkaliphiles and Natronophilic Bacteria. In J. Seckbach, A. Oren & H. Stan-Lotter (Eds.), *Polyextremophiles: Life Under Multiple Forms of Stress* (pp. 123-178). Springer.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics, 30*(15), 2114-2120.

Boros, E., & Kolpakova, M. (2018). A review of the defining chemical properties of soda lakes and pans: An assessment on a large geographical scale of Eurasian inland saline surface waters. *PLoS ONE, 13*(8), e0202205.

Boyd, C. E. (2015). pH, Carbon Dioxide, and Alkalinity. In C. E Boyd (Ed.), *Water Quality* (pp. 153-178). Springer.

Cameron, E. S., Schmidt, P. J., Tremblay, B. J.-M., Emelko, M. B., & Müller, K. M. (2021). Enhancing diversity analysis by repeatedly rarefying next generation sequencing data describing microbial communities. *Scientific Reports, 11*, 22302.

Castelán-Sánchez, H. G., Elorrieta, P., Romoacca, P., Liñan-Torres, A., Sierra, J. L., Vera, I., Batista-García, R. A., Tenorio-Salgado, S., Luzama-Uc, G., Pérez-Rueda, E., Quispe-Ricalde, M. A., & Dávila-Ramos, S. (2019). Intermediate-Salinity Systems at High Altitudes in the Peruvian Andes Unveil a High Diversity and Abundance of Bacteria and Viruses. *Genes, 10*(11), 891.

Chaumeil, P., Mussig, A., Hugenholtz, P., & Parks, D. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics, 36*(6), 1925-1927.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*, i884–i890.

Cisneros-Martínez, A. M., Eguiarte, L. E., & Souza, V. (2023). Metagenomic comparisons reveal a highly diverse and unique viral community in a seasonally fluctuating hypersaline microbial mat. *Microbial Genomics, 9*(7).

Coutinho, F. H., Zaragoza-Salas, A., López-Pérez, M., Barylski, J., Zielezinski, A., Dutilh, B. E., Edwards, R., & Rodriguez-Valera, F. (2021). RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *CellPress, 2*(7).

DasSarma, S., DasSarma, P. (2012). *Halophiles*. John Wiley & Sons.

Dávila-Ramos, S., Castelán-Sánchez, H., Martínez-Ávila, L., Sánchez-Carbente, M., Peralta, R., Hernández-Mendoza, A., Dobson, A. D. W., Gonzalez, R. A., Pastor, N., & Batista-García, R. A. (2019). A Review on Viral Metagenomics in Extreme Environments. *Frontiers In Microbiology, 10*, 2403.

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, F. E., Edwards, R. A., Li, L., Vega-Thurber, R., Reid, R. P., Siefert, J., Souza, V., … Rohwer, F. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature, 452*, 340-345.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M. A., Nelson, K. E., Nilsson, C, Olson, R., Paul, J., Rodriguez-Brito, B., Ruan, Y., Swan, B. K., … Rohwer, F. (2008a). Functional metagenomic profiling of nine biomes. *Nature, 452*, 629-632.

Dinsdale, E. A., Pantos, O., Smriga, S., Edwards, R. A., Angly, F., Wegley, L., Hatay, M., Hall, D., Brown, E., Haynes, M., Krause, L., Sala, E., Sandin, S. A., Bega-Thurber, R., Willis, B. L., Azam, F., Knowlton, N., & Rohwer, F. (2008b). Microbial Ecology of Four Coral Atolls in the Northern Line Islands. *PLoS ONE, 3*, e1584.

Dion, M. B., Plante, P.-L., Zufferey, E., Shah, S. A., Corbeil, J., Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host prediction to decipher the viral dark matter. *Nucleic Acid Research, 49*(6), 3127-3138.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comp. Biol., 7*, e1002195.

Edwards, R. A., McNair, K., Faust, K., Raes, J., Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews, 40*(2), 258-272.

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B., & Banfield, J. F. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol., 78*, 6309–6320.

Escalante, A. E., Eguiarte, L. E., Espinosa-Asuar, L., Forney, L. J., Noguez, A. M., Souza, V. (2008). Diversity of aquatic prokaryotic communities in the Cuatro Cienegas basin. *FEMS Microbiol. Ecol., 65*, 50-60.

Espinosa-Asuar, L., Monroy-Guzmán, C., Madrigal-Trejo, D., Navarro-Miranda, M., Sánchez-Pérez, J., Buenrostro-Muñoz, J., Villar, J., Cifuentes-Camargo, J. F., Kalambokidis, M., Esquivel-Hernandez, D. A., Viladomat-Jasso, M., Escalante, A. E., Velez, P., Figueroa, M., Martinez-Cardenas, A., Ramirez-Barahona, S., Gasca-Pineda, J., Eguiarte, L. E., & Souza, V. (2022). Diversity of an uncommon elastic hypersaline microbial mat along a small-scale transect. *PeerJ, 10*, e13579.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics, 5,* 164-166.

Galiez, C., Siebert, M., Enault, F., Vincent, J., & Söding, J. (2017). WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics, 33*(19), 3113-3114.

García-Ulloa, M., Souza, V., Esquivel-Hernández, D. A., Sánchez-Pérez, J., Espinosa-Asuar, L., Viladomat, M., Marroquín-Rodriguez, M., Navarro-Miranda, M., Ruiz-Padilla, J., Monroy-Guzmán, C., Madrigal-Trejo, D., Rosas-Barrera, M., Vázquez-Rosas-Landa, M., & Eguiarte, L. E. (2022). Recent Differentiation of Aquatic Bacterial Communities in a hydrological System in the Cuatro Ciénegas Basin, After a Natural Perturbation. *Front. Microbiol., 13*, 825167.

Guixa-Boixareu, N., Calderón-Paz, J. I., Heldal, M., Bratbak, G., & Pedrós-Alió, C. (1996). Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. Aquatic *Microbial Ecology, 11*, 215-227.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics, 11*, 119.

Jones, B. E., Grant, W. D., Duckworth, A. W., & Owenson, G. G. (1998). Microbial diversity of soda lakes. *Extremophiles, 2*, 191-200.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from Metagenome Assemblies. *PeerJ, 7*.

Kempe, S., & Kazmierczak, J. (2011). Soda Lakes. In J. Reitner & V. Thiel (Eds.), *Encyclopedia of Geobiology. Encyclopedia of Earth Sciences Series* (pp. 824-828). Springer.

Lanzen, A., Simachew, A., Gessesse, A., Chmolowska, D., Jonassen, I., & Øvreås, L. (2013). Suprising Prokaryotic and Eukaryotic Diversity, Community Structure and Biogeography of Ethiopian Soda Lakes. *PLoS ONE, 8*(8), e72577.

Litchfield, C. D. (2011). Saline Lakes. In J. Reitner & V. Thiel (Eds.), *Encyclopedia of Geobiology. Encyclopedia of Earth Sciences Series* (pp. 765-768). Springer.

López-Lozano, N. E., Eguiarte, L. E., 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*(7), 699-709.

Lu, C., Zhang, Z., Cai, Z., Zhu, Z., Qiu, Y., Wu, A., Jiang, T., Zheng, H., & Peng, Y. (2021). Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol., 19*(5).

Luk, A. W. S., Williams, T. J., Erdmann, S., Papke, R. T., & Cavicchioli, R. (2014). Viruses of Haloarchaea. *Life, 4*, 681-715.

Madrigal-Trejo, D., Sánchez-Pérez, J., Espinosa-Asuar, L., Valdivia-Anistro, J. A., Eguiarte, L. E., & Souza, V. (2023). A metagenomic time-series approach to assess the ecological stability of microbial mats in a seasonally fluctuating environment. *Microbial Ecology*.

Madrigal-Trejo, D. (2022). *Análisis del metametaloma en los tapetes microbianos de Domos del Arqueano, Cuatro Ciénegas, como recapitulación del uso de metales a lo largo de la historia de la Tierra* [Bachelor's dissertation, Universidad Nacional Autónoma de México].

McDaniel, L., Breibart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., & Paul, J. H. (2008). Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS ONE, 3*, e3263.

McGenity, T. J., & Oren, A. (2012). Hypersaline Environments. In E. M. Bell, (Ed.), *Life at Extremes: Environments, Organisms and Strategies of Survival* (pp. 402-437). CAB International.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE, 8*, e61217.

Medina, L. E., Taylor, C. D., Pachiadaki, M. G., Henríquez-Castillo, C., Ulloa, O., & Edgcomb, V. P. (2017). A review of protist grazing below the photic zone emphasizing studies of oxygen-depleted wáter columns and recent applications of *in situ* approaches. *Front. Mar. Sci., 4*, 105.

Medina-Chávez, N. O., Viladomat-Jasso, M., Zarza, E., Islas-Robles, A., Valdivia-Anistro, J., Thalasso-Siret, F., Eguiarte, L. E., Olmedo-Álvarez, G., Souza, V., & De la Torre-Zavala, S. (2023). A transiently hypersaline microbial mat harbors a diverse and stable archaeal community in the Cuatro Cienegas Basin, Mexico. *Astrobiology, 8*.

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications, 7*, 11257.

Merino, N., Aronson, H. S., Bojanova, D. P., Feyhl-Buska, J., Wong, M. L., Zhang, S., & Giovannelli, D. (2019). Living at Extremes: Extremophiles and the Limits of Life in a Planetary Context. *Front. Microbiol., 10*, 780.

Minegishi, H. (2013). Halophilic, Acidophilic and Haloacidophilic Prokaryotes. In J. Seckbach, A. Oren & H. Stan-Lotter (Eds.), *Polyextremophiles: Life Under Multiple Forms of Stress* (pp. 203-213). Springer.

Moreno-Letelier, A., Olmedo-Alvarez, G., Eguiarte, L. E., & Souza, V. (2012). Divergence and Phylogeny of Firmicutes from the Cuatro Ciénegas Basin, Mexico: A Window to an Ancient Ocean. *Astrobiology, 12*(7), 674-684.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research, 27*(5), 824-834.

Parks, D., Imelfort, M., Skennerton, C., Hugenholtz, P., & Tyson, G. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25*(7), 1043-1055.

Prangishvili, D., Bamford, D. H., Forterre, P., Iranzo, J., Koonin, E. V., & Krupovic, M. (2017). The enigmatic archaea virosphere. *Nat. Rev. Microbiol., 15*(12), 724-739.

Purdy, K. J. (2005). Nucleic acid recovery from complex environmental samples. *Methods Enzymol., 397*, 271–292.

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Available from: hhttps://www.R-project.org/

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics, 16*(6), 276-277.

Rocke, E., Pachiadaki, M. G., Cobban, A., Kujawinski, E. B., & Edgcomb, V. P. (2015). Protist community grazing on prokaryotic prey in deep ocean water masses, *PLoS ONE, 10*(4), e0124505.

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.

Roux, S., Enault, F., Ravet, V., Colombet, J., Bettarel, Y., Auguet, J.-C., Bouvier, T., Lucas-Staat, S., Vellet, A., Prangishvili, D., Forterre, P., Debroas, D., & Sime-Ngando, T. (2016). Analysis of metagenomic data reveals common features of halophilic viral communities across continents. *Environmental Microbiology, 18*(3), 889-903.

Roux, S., Camargo, A. P, Coutinho, F. H., Dabdoub, S. M., Dutilh, B. E., Nayfach, S., & Tritt, A. (2023). iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.*, *21*(4), e3002083.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020). CRISPRCASTYPER: Automated Identification, annotation, and classification of CRISPR-Cas Loci. *The CRISPR Journal, 3*(6), 462–469.

Saccò, M., White, N. E., Harrod, C., Salazar, G., Aguilar, P., Cubillos, C. F., Meredith, K., Baxter, B. K., Oren, A., Anufriieva, E., Shadrin, N., Marambio-Alfaro, Y., Bravo-Naranjo, V., & Allentoft, M. E. (2021). Salt to conserve: a review on the ecology and preservation of hypersaline ecosystems. *Biol. Rev*., 000-000.

Santos, F., Meyerdierks, A., Peña, A., Rosselló-Mora, R., Amann, R., & Antón, J. (2007). Metagenomic approach to the study of halophages: the environmental halphage 1. *Environmental Microbiology, 9*(7), 1711-1723.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecol. Lett., 10*, 1115–1123.

Song, W., & Thomas, T. (2017). Binning_refiner: improving genome bins through the combination of different binning programs. *Bioinformatics, 3*3(12), 1873-1875.

Souza, V., Espinosa-Asuar, L., Escalante, A. E., Eguiarte, L. E., Farmer, J., Forney, L., lloret, L., Rodríguez-Martínez, J. M., Soberón, X., Dirzo, R., & Elser, J. J. (2006). An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *PNAS, 103*(17), 6565-6570.

Souza, V., Siefert, J. L., Escalante, A. E., Elser, J. J., Eguiarte, L. E. (2012). The Cuatro Ciénegas Basin in Coahuila Mexico: An Astrobiological Precambrian Park, *Astrobiology, 12*(7), 641-647.

Souza, V., Moreno-Letelier, A., Travisano, M., Alcaraz, L. D., Olmedo, G., & Eguiarte, L. E. (2018). The lost world of Cuatro Ciénegas Basin, a relictual bacterial niche in a desert oasis. *eLife, 7*, e38278.

Sullivan, M. B., Weitz, J. S., & Wilhelm, S. (2017). Viral ecology comes of age. *Environmental Microbiology Reports, 9*, 33-35.

Taboada, B., Isa, P., Gutiérrez-Escolano, A. L., del Ángel, R. M., Ludert, J. E., Vázquez, N., Tapia-Palacios, M. A., Chávez, P., Garrido, E., Espinosa, A. C., Eguiarte, L. E., López, S., Souza, V., & Arias, C. F. (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.

Turner, D., Shkoporov A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., van Zyl, L. J., Aziz, R. K., Oksanen, H. M., Poranen, M. M., Kropinski, A. M., Barylski, J., Brister, J. R., Chanisvili, N., Edwards, R. A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., … Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcomittee. *Archives of Virology, 168*, 74.

Vavourakis, C. D., Ghai, R., Rodriguez-Valera, F., Sorokin, D. Y., Tringe, S. G., Hugenholtz, P., & Muyzer, G. (2016). Metagenomic Insights into the Uncultured Diversity and Physiology of Microbes in Four Hypersaline Soda Lake Brines. *Front. Microbiol., 7*(211).

Ventosa, A., Fernández, A. B., León, M. J., Sánchez-Porro, C., & Rodriguez-Valera, F. (2014). The Santa Pola saltern as a model for studying the microbiota of hypersaline environments. *Extremophiles, 18*, 811–824

Villarroel, J., Kleinheinz, K. A., Jurtz, V. I., Zschach, H., Lund, O., Nielsen, M., & Larze, M. V. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses, 8*(5), 116.

Walsh, M. M. (2010). Microbial Mats on the Early Earth: The Archaean Rock Record. In J. Seckbach & A. Oren (Eds.), *Microbial Mats: Modern and Ancient Microorganisms in Stratified Ecosystems* (pp. 43-51). Springer.

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., Braun, J., Sun, F., & Ahlgre, N. A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics, 2*(2), lqaa044.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Winter, C., Bouvier, T., Weinbauer, M. G., & Thingstad, T. F. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the “Killing the Winner” hypothesis revisited. *Microbiology and Molecular Biology Reviews, 74*, 42-57.

Wolaver, B. D., Crossey, L. J., Karlstrom, K. E., Banner, J. L., Cardenas, M. B., Gutiérrez-Ojeda, C., & Sharp, J. M. (2013). Identifying origins of and pathways for spring waters in a semiarid basin using He, Sr, and C isotopes: Cuatrociénegas Basin, Mexico. *Geosphere, 9*(1), 113-125.

Wommack, K. E., & Colwell, R. R. (2000). Viroplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews, 64*(1), 69-114.

Wu, Y., Simmons, B., & Singer, S. (2015). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics, 32*(4), 605-607.

Yachi, S., & Loreau, M. (1999). Biodiversity and ecosystem productivity in a fluctuating environment: The insurance hypothesis. *PNAS, 96*, 1463-1468.

Zhou, F., Gan, R., Zhang, F., Ren, C., Yu, L., Si, Y., & Huang, Z. (2022). PHISDetector: A tool to detect divrerse in silico phage-host interaction signals for virome studies. *Genomics, Proteomics & Bioinformatics, 20*(3), 508-523.