UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS QUÍMICAS Y FARMACEÚTICAS

COMUNIDADES VIRALES DEL CAMPO GEOTERMAL EL TATIO Y OTROS SISTEMAS TERMALES TERRESTRES DEL MUNDO

Tesis presentada a la Universidad de Chile para optar al grado de Magíster en Bioquímica, área de Especialización en Bioquímica Ambiental y Memoria para optar al Título de Bioquímico por:

FELIPE ANDRÉS LOYOLA AHUMADA

Director(es) de tesis: Dra. Beatriz Díez Moreno
Dr. Davor Cotoras Tadic

Santiago-Chile
Marzo 2024

UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS QUÍMICAS Y FARMACÉUTICAS

INFORME DE APROBACIÓN DE TESIS DE MAGÍSTER

Se informa a la Dirección de la Escuela de Graduados de la Facultad de Ciencias Químicas y Farmacéuticas que la Tesis de Magíster y Memoria de Título presentada por el candidato

FELIPE ANDRÉS LOYOLA AHUMADA

Ha sido aprobada por la Comisión de Evaluadora de Tesis como requisito para optar al grado de Magíster en Bioquímica, Área de Especialización: Bioquímica Ambiental y Título de Bioquímica, en el examen público rendido el día _____

Director de Tesis:
Dra. Beatriz Díez _____

Director de Tesis:
Dr. Davor Cotoras _____

Comisión Evaluadora de Tesis:
Dra. Julieta Orlando _____

Dr. Vinicius Maracaja _____

Dr. Sergio Álvarez _____

Dedicatoria

A mi madre Carola y mi padre Gustavo,

por su infinito apoyo incondicional

y confianza en mí.

Agradecimientos

Primero a mi familia, mi madre y padre que estuvieron siempre que necesité ayuda, sin ellos nada de esto habría sido posible, muchas gracias por su dedicación conmigo.

A mi tutora Beatriz, quién me ayudó en cada etapa de este proceso. Muchas gracias por todas las oportunidades que me brindó y la confianza que depositó en mí. Profe, usted es una increíble persona y científica, la forma en que disfruta la ciencia y transmite su pasión, tan noblemente, realmente inspira y para mí fue un motor que me permitió volver a enamorarme de la ciencia.

A todo el querido Lab BD, que me apoyaron inmensamente durante todo el trayecto, Oscar, Sergio, Marianne, Pablo, Jaime, Carla, Javier, Karen, Christina, Natali, Tomás, Johanna, JB, Felipe, Bárbara y especialmente Wilson, sin tu constante ayuda, guía, consejo y apoyo cuando las cosas se veían complejas, este proceso habría sido mucho más difícil y menos ameno, eres el mejor.

A mis mejores amigos de la universidad, Andrea, Daniel, Diego, Jorge, Loreto y Wilson, además del equipo de balonmano de la facultad, hicieron que todos estos años fueran de los mejores de mi vida hasta la fecha.

A los Gatitos Bonitos y a los P.B por ser un núcleo de apoyo antes, durante y después de esta etapa de mi vida.

En fin, ¡¡gracias, gracias y más gracias!!

Presentaciones a Congresos

Elucidating Viral Communities in El Tatio Geysers Microbial Mats. **Felipe Loyola**, Oscar Salgado, Sergio Guajardo, Beatriz Díez (2021). Association for the Sciences of Limnology and Oceanography (ASLO) 2021 Aquatic Sciences Virtual Meeting. 22-27 de Junio 2021.

Viral Communities of Terrestrial Thermal Systems of the World and Their Ecological Drivers. **Felipe Loyola-Ahumada**, Sergio Guajardo-Leiva, Oscar Salgado, Beatriz Díez (2022). XLIV Congreso Chileno de Microbiología. La Serena, Chile. 29 de noviembre al 02 de diciembre 2022.

Viral Communities of Terrestrial Hot Springs Worldwide: Biogeography and Ecological Features. **Felipe Loyola**, Sergio Guajardo-Leiva, Oscar Salgado, Beatriz Díez (2023). 3er. Congreso Latinoamericano de Ecología Microbiana, ISME-Lat 2023. Universidad Nacional de Quilmes, Bernal, Buenos Aires, Argentina. 07-10 de agosto 2023.

Índice

Índice de figuras

Índice de figuras suplementarias

Índice de tablas suplementarias

Resumen

Los virus han demostrado ser ubicuos en todos los ambientes y las aguas termales no son la excepción, a pesar de sus condiciones extremas. En su mayoría, los virus termales como los de otros ambientes son desconocidos, ya que la ausencia de biomarcadores conservados universalmente dificulta su identificación. Una solución ha sido utilizar la meta-ómica. Sin embargo, son muy pocos los estudios que buscan caracterizar la composición y estructura de las comunidades virales en estos ambientes extremos, así como identificar los factores ambientales que las determinan y su biogeografía. En este estudio, el objetivo fue caracterizar a nivel global las comunidades virales de sistemas termales, analizando un total de 49 metagenomas de cuatro continentes (siete países), mediante múltiples herramientas bioinformáticas. Este análisis generó un total de 3.559 unidades taxonómicas operativas virales (vOTUs) que se utilizaron para establecer redes de genes compartidos y estudios de diversidad. La mayoría de las comunidades mostró afiliación cercana a la clase *Caudoviricetes*, no obstante, más del 50% no logró ser asignado, indicando una alta novedad taxonómica. Los resultados de este estudio también revelaron que las comunidades virales son afectadas por factores ecológicos, siendo la distancia geográfica el principal determinante de la distribución diferencial de vOTUs, interactuando a nivel local con factores fisicoquímicos. Este estudio, por tanto, reveló patrones biogeográficos a nivel de comunidad que sugieren una dispersión global limitada, mientras una circulación y adaptación más local.

Summary

Viruses have proven to be ubiquitous in all environments on earth and hot springs are not the exception, despite their extreme conditions. Most of the viruses, independently of its environment, are unknown, by the lack of universally conserved viral genes, which hampers its identification.

One way to address this problem is through meta omics. However, very few studies have sought to characterize their presence, composition, and structure in thermophilic geothermal springs, as well as to identify the ecological drivers that determine them and their biogeography.

In this study, the objective was to globally characterize the viral communities of hot springs by analyzing a total of 49 metagenomes from four continents (seven countries) using multiple viral identification and quality analysis tools. This analysis generated a total of 3,559 viral operational taxonomic units (vOTUs) that were used to establish shared gene networks and diversity studies.

Taxonomically, most communities globally showed close affiliation to the order Caudovirales, with the family Myoviridae predominating, however, more than 50% could not be assigned, indicating high taxonomic novelty. The results of this study also revealed that viral communities are affected by ecological factors, with geographical distance being the main determinant of differential vOTUs distribution, interacting locally with physicochemical factors. This study, therefore, revealed biogeographical patterns at the viral community level that suggest limited global dispersal, while a more local circulation and adaptation.

1. Introducción

Los organismos procariotas y sus virus asociados están presentes en todos los ambientes del planeta, alterando de manera significativa la biosfera y sus procesos debido al rol que juegan en los ciclos biogeoquímicos y de nutrientes (Howard-Varona et al., 2017). En estas interacciones, los virus juegan un rol integral en regular a las comunidades microbianas y mantener el funcionamiento de los ecosistemas, por ejemplo, mediante interacciones virus- hospedero que facilitan la transferencia horizontal de genes entre comunidades microbianas y contribuyen a la renovación de los nutrientes mediante lisis celular del hospedero (Zablocki et al., 2018).

1.1 Estado de la investigación viral

En la actualidad, para comprender la dinámica virus-hospedero se sigue el modelo Lotka-Volterra, conocido coloquialmente como Kill-The-Winner (KTW), el cual predice que los virus reducen rápida y drásticamente la población de las especies microbianas más abundantes, previniendo así que los mejores competidores microbianos generen una alta biomasa (Rodriguez-Brito et al., 2010). Sin embargo, una barrera para comprender la dinámica de los virus y su rol específico a nivel de ecosistema es que, a diferencia de las bacterias que poseen ARN ribosomal que es utilizado como genes marcadores universales, no existen genes universales o métodos formalizados que permitan una taxonomía o análisis comparativo (Jang et al., 2019). Esto, por tanto, restringe los análisis de diversidad viral independientes de cultivo, siendo, además, difícil cultivar virus dada la falta de hospederos cultivables a los que infecten. Para evitar estas

limitaciones, se ha estado utilizando la secuenciación de escopeta de metagenomas virales (o metaviromas) para permitir la secuenciación extensa del ADN (ácido desoxirribonucleico) de una variedad de partículas virales presentes en el medioambiente (Moon et al., 2021), las cuales, en conjunto con la secuenciación del ARN (ácido ribonucleico) (metatranscriptomas), también han permitido evaluar la actividad que presentan estos viriones (New et al., 2020). Gracias a esto, nuestro conocimiento sobre los virus está rápidamente aumentando al aplicar los avances realizados en métodos computacionales para la detección de secuencias virales de ADN/ARN, permitiendo el desarrollo de grandes bases de datos de genomas virales completos y proteínas virales (Jarett et al., 2020).

1.2 Taxonomía de virus

Dada la velocidad con la que se está expandiendo el conocimiento respecto de los virus, se ha vuelto urgente la necesidad de establecer una taxonomía viral basada en genomas que nos permita una red de clasificación global. Para esto han surgido diversas estrategias para desarrollar redes taxonómicas para virus de bacterias y arqueas (fagos) o eucariotas. Para los virus bacterianos/arqueanos o fagos, metodologías más antiguas utilizaron la comparación de secuencias de proteínas emparejadas con genomas completos en redes filogenéticas concordantes con la base de datos del Comité Internacional de Taxonomía de Virus (ICTV) (Lefkowitz et al., 2018). Sin embargo, este acercamiento no pudo ser completamente avalado debido al mosaicismo generalizado de los fagos, donde diferentes regiones de estos

genomas tienen historias evolutivas distintas, debido a la transferencia horizontal de genes (Filée et al., 2006; Dion et al., 2020), que podría enturbiar los límites taxonómicos y romper las suposiciones bajo la cual trabajan los algoritmos. Otros acercamientos han estimado la fracción de genes compartidos y el porcentaje de identidad de dichos genes para definir las afiliaciones de géneros y subfamilias, pero este acercamiento también falla en definir la clasificación taxonómica para muchos grupos virales debido a la cantidad de posibilidades que representa la alta variabilidad evolutiva de los virus procariotas (Lavigne et al., 2008, 2009; Mavrich y Hatfull, 2017). A pesar de esto, redes de intercambio de genes basadas en grupos de proteínas entre genomas virales ha demostrado ser ampliamente concordante con los taxones del ICTV, independiente de si se utilizan redes monopartitas (es decir, sólo genomas virales) o bipartitas (genomas virales y genes) para grandes conjuntos de datos provenientes de, por ejemplo, metagenomas (Jang et al., 2019).

1.3 Virus en el medio ambiente

Proyectos internacionales multidisciplinarios como las expediciones del "*Tara* Oceans" han evaluado, entre otros, la complejidad de la vida oceánica viral a través de una escala taxonómica y espacial profunda, ilustrando exitosamente cómo los conceptos y datos a escala global pueden ayudar a integrar la complejidad biológica en modelos y servir como una línea de base para evaluar los cambios en ecosistemas (Sunagawa et al., 2020).

Estos estudios ecológicos para las comunidades virales a grandes escalas se han realizado principalmente en los ecosistemas marinos, debido a la influencia

de este sistema en los ciclos biogeoquímicos, considerando, por ejemplo, que los microorganismos oceánicos producen la mitad del oxígeno que respiramos (Field et al., 1998), donde los virus marinos son presumiblemente actores clave en estas interacciones (Suttle, 2007; Breitbart, 2012; Brum et al., 2015).

1.4 Comunidades microbianas en sistemas termales

Existen otros ecosistemas dinámicos, pero relativamente más simples, donde la complejidad de las comunidades microbianas puede ser menor, éstos son los ambientes termales, los que han ayudado a entender y poner a prueba principios de ecología microbiana (Ward et al., 1998; Ward, 2006; Bhaya et al., 2007; Klatt et al., 2011; Alcamán et al., 2018; Guajardo-Leiva et al., 2021).

Uno de los beneficios de estudiar comunidades virales en los hábitats geotermales reside en el hecho de que las comunidades microbianas a las que infectan los virus en estos ambientes a altas temperaturas son generalmente menos diversas que en suelos, aguas marinas, sedimentos o lagunas, y, por lo tanto, ofrecen una gran oportunidad para estudiar la estructura y función de diferentes comunidades virales utilizando la metagenómica ambiental (Inskeep et al., 2013). Además, estos ambientes han servido por décadas como modelos para probar el concepto de endemismo en microorganismos, al considerarse como islas biogeográficas (Papke et al., 2003), donde el aislamiento físico y geográfico nos permite explorar la generación de divergencia genética y especiación alopátrica entre microorganismos (Papke et al., 2003, Whitaker et al., 2003, Martiny et al., 2006). Como ejemplo, varios estudios han encontrado la existencia de filotipos de la cianobacteria *Synechococcus* en termas del Tíbet y

otros continentes, únicas de cada lugar a pesar de la similitud de variables geoquímicas entre termas, lo que apoya la noción de que existen efectos de distancia geográfica a una escala intercontinental en estas poblaciones, y por tanto evidencia de la existencia de una biogeografía microbiana (Lau et al., 2009).

1.5 Tapetes microbianos de sistemas termales

Estos ambientes extremos para la vida, típicamente de estructura microbiana de baja complejidad están dominados por microorganismos que forman tapetes bien definidos que son constantemente recorridos por aguas termales (Alcamán et al., 2015) y que abarcan un amplio rango de propiedades fisicoquímicas que surgen de varios procesos geológicos. Sus temperaturas varían entre 40°C y 98°C, siendo clasificados los microorganismos que los habitan como moderadamente termofílicos (40-71°C) o hipertermofílicos (72-98°C). Los valores de pH de estos ambientes van desde 1 a 9, definidos en 3 categorías, ácidos (pH 1-5), neutros (pH 6-7,5) y alcalinos (pH > 7,5) (Zablocki et al., 2018). La temperatura ha sido considerada una de las variables más importantes asociadas a cambios y adaptaciones metabólicas en las comunidades presentes en los tapetes microbianos en sistemas termales con pH neutro (Klatt et al., 2011; Mackenzie et al., 2013; Alcorta et al., 2020; Bennet et al., 2022). Frecuentemente, la capa externa de estos tapetes está compuesta por organismos fotoautótrofos oxigénicos, tales como cianobacterias, incluyendo *Synechococcus* spp., y otras cianobacterias filamentosas sin heterocistos, como *Oscillatoria* spp., o con heterocistos, como *Fischerella* spp., así como también fotótrofos filamentosos anoxigénicos (FAPs), tales como *Roseiflexus* sp. y *Chloroflexus* sp. (Alcamán et

al 2018., Guajardo-Leiva et al., 2018; Kees et al., 2022; Hamilton y Havig, 2022; George et al., 2023; Moreno et al., 2023). Varios estudios de diversidad en tapetes microbianos de ambientes termales han mostrado que miembros de los filos Cyanobacteria y Chloroflexi pueden coexistir, a veces incluso de manera colaborativa (Miller et al., 2009; Alcamán et al., 2018; Bennet et al., 2020). Así, estas simplificadas, pero altamente cooperativas comunidades son de gran utilidad para comprender la composición, estructura y función de los consorcios microbianos (Klatt et al., 2011), e investigar el rol de factores abióticos, tales como el pH, concentración de sulfuros y temperaturas, que determinan las comunidades y ciclos biogeoquímicos en estos ecosistemas. Sin embargo, hay una falta de investigación en cuanto a los factores bióticos, como los virus en estos tapetes termofílicos fotoautotróficos (Guajardo-Leiva et al., 2018, Zablocki et al., 2018).

1.6 Virus en tapetes microbianos de ecosistemas termales

En estos ambientes termales terrestres, el estudio ha estado principalmente enfocado en el descubrimiento y caracterización de virus de arqueas (Jarret et al., 2020), mientras las comunidades virales en estos ambientes también comprenden virus de bacterias termofílicas y probablemente virus de eucariotas (virus de algas y micovirus). Si bien se han investigado ciertos aspectos del impacto ecológico y abundancia de los fagos presentes en estos ambientes extremos, siguen siendo poco estudiados, por lo que se cuenta con pocos genomas secuenciados (en comparación a otros ecosistemas) (Jarret et al, 2020), además de que dichos estudios microbianos se han dirigido

principalmente a las fuentes de agua termal (Zablocki et al., 2017). En estas aguas los virus abundan en rangos que van desde 10^4 a 10^9 partículas similar a virus (VLP, por sus siglas en inglés) mL^{-1}, jugando un rol importante en la estructura de las poblaciones de hospederos y como motor de reciclaje de nutrientes orgánicos e inorgánicos (Guajardo-Leiva et al., 2018). A pesar de los pocos estudios metavirómicos de estos sistemas acuáticos, estos indican que los virus termófilos ambientales difieren de los obtenidos en cultivos, encontrándose solamente entre 20 y 50% de similitud entre las secuencias virales ambientales obtenidas y las presentes en las bases de datos (Bolduc et al., 2015). Hasta ahora, los genomas de virus termofílicos que han sido aislados y secuenciados (57 genomas, de los cuales 37 infectan arqueas y 20 bacterias) mostraron pocas coincidencias con secuencias de las bases de datos públicas. Esto, sumado a evidencias de tapetes microbianos que en gradientes de temperatura entre 46 y 71°C están dominados por fotótrofos bacterianos como las cianobacterias (predominando sobre arqueas), sugiere que hay una gran cantidad de fagos desconocido en estos ambientes (Guajardo- Leiva et al., 2018). Así, en la actualidad, contamos con un conocimiento muy limitado sobre la diversidad viral en tapetes microbianos de termas, reduciéndose a pocos sitios en el mundo. Estudios como el de Zablocki et al. en las termas de Brandvlei, Sudáfrica (primera terma africana que se estudia) han revelado la presencia de Caudovirales y genomas asociados a Podoviridae, encontrándose que el “cóntigo” (derivado de la palabra “contiguo”, es un conjunto de secuencias de ADN que se superponen de forma que colectivamente dan una representación continua de una región

genómica) (Gregory, 2005; Gibson y Muse., 2009) viral más representativo correspondía a un fragmento del genoma de un cianofago (Zablocki et al., 2017). Trabajos realizados en nuestro laboratorio en la terma Porcelana en la Patagonia chilena, también han revelado que la mayoría de la comunidad viral pertenece al orden de los Caudovirales (70% del total de secuencias virales), siendo la mayor cantidad de actividad infectiva representada por cianofagos (Guajardo-Leiva et al., 2018). Además, trabajos no publicados de nuestro grupo también muestran la presencia de Caudovirales en dos muestras de tapetes microbianos fotoautótrofos del campo geotermal chileno El Tatio a diferentes temperaturas. En estos tapetes, al menos el 50% de las secuencias virales no estaban relacionadas a ningún virus previamente secuenciado y los que estaban relacionados a familias Caudovirales conocidas (Myoviridae, Siphoviridae y Podoviridae) representan nuevos genomas de virus reportados de aguas termales, sugiriendo que la mayoría de la comunidad viral de El Tatio podría ser específica para este entorno. Además, la taxonomía de los virus ha experimentado significativos cambios últimamente, reclasificando, entre otros, a los bacteriófagos con cola de genoma de doble hebra de ADN, donde el conocido orden *Caudovirales*, que incluía a las familias *Myoviridae*, *Siphoviridae* y *Podoviridae*, fue abolido por el Comité Internacional de Taxonomía de Virus (Lefkowitz et al., 2018). Estos fagos bacterianos están siendo reubicados junto a los virus con morfología de cola y cabeza que infectan arqueas (Evseev et al., 2023) y herpesvirus en el reino *Heunggongvirae*, atribuyendo a los bacteriófagos al filo *Uroviricota* y la clase *Caudoviricetes* (Turner et al., 2023).

Con estos antecedentes y razones, es que en el presente proyecto también se pretende extender los estudios meta-ómicos del campo geotermal de El Tatio a través del análisis de 13 termas más, y evaluar la identidad, diversidad y actividad de estas comunidades de fagos. Este estudio representa uno de los trabajos más grandes de ecología viral conocidos en ambientes termales terrestres a escala local y regional. Este trabajo, esperamos que ayude finalmente a catalogar la biodiversidad existente bajo la particular condición geoquímica de El Tatio, que a su vez comparte características con otras termas alrededor del mundo, permitiendo comparaciones globales.

2. Hipótesis

H1. “Las comunidades virales en los tapetes microbianos de las termas de El Tatio están estructuradas de forma diferente en función de los factores fisicoquímicos y la distancia espacial que las separa, pero comparten los mismos hospederos”.

H2. “Las comunidades virales en los tapetes microbianos de las termas de El Tatio presentan una alta novedad taxonómica, con miembros únicos y distintos a los de otros sistemas termales del mundo debido a la distancia geográfica y aislamiento espacial de El Tatio”.

3. Objetivo General

Determinar la composición, diversidad, actividad y hospederos de las comunidades virales de tapetes microbianos en el campo geotermal El Tatio, así como comparar su estructura con otros sistemas termales del mundo de distintas ubicaciones geográficas en un rango más amplio de los factores fisicoquímicos.

4. Objetivos Específicos

1. Determinar la composición y diversidad de las comunidades virales de ADN presentes en tapetes microbianos de termas en El Tatio.
2. Caracterizar las comunidades virales activas en tapetes microbianos de termas de El Tatio para determinar los principales hospederos.
3. Comparar la composición y diversidad de las comunidades de fagos de El Tatio con las de otras fuentes termales en un mayor rango de los factores fisicoquímicos y con distintas ubicaciones geográficas en el mundo.

5. Metodología

5.1 Sitios de muestreo de El Tatio

El campo de géiseres El Tatio está ubicado en los montes andinos del norte de Chile, en la región de Antofagasta (22°19′53″S 68°0′37″O), a unos 4200 metros sobre el nivel del mar. Las muestras colectadas en varias termas de este campo en enero del año 2020 presentan un pH relativamente neutro que oscila entre 6,8 a 8,6 (con la excepción de un punto de muestreo que presenta un pH de 9,27) y un rango de temperaturas que van desde los 45°C a 62°C. Se obtuvieron 16 muestras de tapete microbiano, pertenecientes a 13 puntos (termas) escogidos estratégicamente para cubrir el campo geotermal en su completitud. La temperatura de las muestras se midió con una cámara infrarroja de barrido frontal (Fluke TiS45, WA, USA) y se corroboró en el tapete con un instrumento multiparámetro (WTW multi 340i, NY, USA) que además midió el pH. Se tomaron muestras en triplicado de aproximadamente 2 mililitros (mL) mediante una perforación del tapete utilizando un sacabocado para cada punto de muestreo seleccionado. Las muestras se mantuvieron en viales criogénicos con RNALater (Thermo Fischer Scientific, Vilnius, Lithuania) y luego se almacenaron a -80 °C hasta que se les hizo una extracción de ADN.

5.2 Extracción de ADN y secuenciación de tapetes microbianos

El ADN fue extraído utilizando un protocolo modificado de Tillet y Neillan (2000). Brevemente, se ubican las muestras en tubos (E) de matriz de lisis (Qbiogene, Carlsbad, CA, USA) que contienen un buffer XS (1% Etil Xantogenato potásico, una concentración 100 milimolar (mM) de Tris-HCl, pH 7,4; 20 mM EDTA, pH 8;

1% dodecil sulfato de sodio; 800 mM acetato de amonio) y cuentas de vidrio de 1 milímetro (mm) para obtener un lisado de células por agitación de cuentas (BeadBeater) en dos ciclos de 1 minuto a 4350 revoluciones por minuto (rpm). Luego, 10% de SDS fue agregado y las muestras se incubaron durante dos horas a 65°C. Las muestras se agitaron mediante vórtex por unos segundos, se incubaron en hielo por 30 minutos y se centrifugaron a máxima velocidad durante 10 minutos a 4°C. Después, se hacen dos pasos de extracción utilizando fenol, cloroformo y alcohol isoamilo (IAA) en una proporción 25:24:1 (fenol:cloroformo:IAA) y uno con cloroformo:IAA (25:1). Los ácidos nucleicos se precipitaron a -20°C por dos horas con isopropanol absoluto junto a un 1/10 volumen de acetato de amonio 4M. El pellet fue lavado con etanol al 70% y finalmente resuspendido en 50-100 microlitros (µL) de agua libre de nucleasas. Cantidades mayores a 100 ng de ADN por muestra fueron almacenadas y enviado en tubos DNAstable (Biomatrica, San Diego, CA, USA) al Centro Biotecnológico Roy J. Carver (Universidad de Illinois, Urbana-Champaign, IL, USA), donde se hicieron librerías utilizando KAPA HyperPrep (Kapa Biosystems, Roche, Basel, Switzerland), para ser luego secuenciadas en la plataforma Illumina Novaseq 6000 (celdas de flujo S1, 2 x 150 pb). Una vez recibidas las secuencias, se realizó un filtrado de calidad a las lecturas crudas de acuerdo con Guajardo-Leiva et al. (2018). Brevemente, se aplicaron varios filtros utilizando la herramienta Cutadapt (Martin, 2011), dejando sólo secuencias mapeables de más de 30 pb (-m 30) con un recorte en el extremo 3’ con una calidad menor a 28 (-q 28), así como la eliminación y no alineación de las primeras 5 bases a la

izquierda (-u 5) y una coincidencia perfecta de al menos 10 pb (-O 10) frente a un adaptador Illumina estándar. Finalmente, se eliminaron las secuencias que representan repeticiones simples (usualmente por errores de secuenciación) con el programa PRINSEQ (Schmieder y Edwards, 2011) modo DUST (-lc_method dust, -lc_threshold 7).

5.3 Extracción de ARN y secuenciación de tapetes microbianos en El Tatio.

En el caso de las muestras de ARN, se realizó una extracción con Trizol a partir del mismo tipo de muestras mencionadas previamente (tapete con RNALater). Brevemente, se partió con un lavado de la muestra de tapete (para eliminar el RNALater) mediante centrifugaciones sucesivas con agua bidestilada (1 ml, tres veces), para luego ser homogenizado con cuentas de vidrio y TRIzol® (Invitrogen) en un BeadBeater y finalmente se procedió con la extracción de ARN. El ARN obtenido se purificó y concentró utilizando el kit RNA Clean & Concentrator (Zymo Research, USA). La secuenciación se realizó en el centro biotecnológico Roy J. Carver (Universidad de Illinois, Urbana-Champaign, IL, USA), con el sistema de secuenciación Illumina Novaseq 6000.

5.4 Metagenomas de sistemas termales del mundo

En este estudio, se definió un rango de temperatura termofílico entre 40°C y 80°C y un pH aproximadamente neutral (6-9) como los principales factores fisicoquímicos para elegir los sistemas termales para realizar los análisis comparativos. Estos parámetros excluyen a la mayoría de procariontes acidófilos y arqueas hipertermofílicas, mientras que permiten la predominancia de tapetes

microbianos fototróficos (Alcamán-Arias et al., 2018; Salgado et al., 2022). En total se analizaron 49 metagenomas (Tabla suplementaria 1), de los cuales 32 corresponden a datos disponibles públicamente que representan 3 continentes (Asia, América y Antártica) y 7 países (Canadá, Chile, China, Colombia, India, Japón y Estados Unidos), mientras que los 17 metagenomas restantes utilizados corresponden a muestreos realizados por nuestro grupo de investigación en el campo geotermal de géiseres de El Tatio.

5.5 Ensamblaje de metagenomas

Se realizó un ensamblaje *de novo* para los 49 metagenomas utilizando el programa SPAdes v3.10.1 (-meta) (Bankevich et al., 2012). Uno de los resultados de este ensamble corresponde a los “cóntigos” (o contigs, como se les llamará de aquí en adelante) que son una representación de un set de segmentos de ADN o secuencias que se superponen de manera que proporcionan una representación contigua de una región genómica (Gregory, 2005; Gibson y Muse, 2009).

5.6 Minado viral

La búsqueda de secuencias virales en los 49 metagenomas fue realizada con un enfoque multiherramienta utilizando la interfaz “What the Phage” (WtP) v1.0.1 (Marquet et al., 2022). Esta metodología consiste en un flujo de trabajo que aúna múltiples herramientas bioinformáticas para la identificación de fagos, que, en conjunto con otras estrategias de anotación y clasificación, facilitan al usuario la toma de decisiones cuando las distintas herramientas no están en acuerdo y así seleccionar los contigs virales más probables para siguientes análisis más detallados (Marquet et al., 2022). Una vez que se obtuvieron los contigs del ensamblado metagenómico, se realiza un filtro por tamaño de secuencias, donde se seleccionan solamente las secuencias que superen un umbral de 10.000 pb (Roux et al., 2017). Esto se realiza utilizando la herramienta “awk” (lenguaje diseñado para procesar datos basados en texto) del sistema operativo UNIX (Dougherty y Robbins., 1990). Este filtro busca eliminar las secuencias que son muy pequeñas, ya que usualmente generan falsos positivos (Gregory et al., 2019). La predicción de fagos la realizan 11 herramientas distintas en paralelo: VirFinder v1.1 (Ren et al., 2017), PPR-Meta v1.1 (Fang et al., 2019), VirSorter v1.0.6 (Roux et al., 2015), DeepVirFinder v1.0 (Ren et al., 2020), Metaphinder (Jurtz et al., 2016), sourmash v2.0.1 (Brown et al., 2016), Vibrant v1.2.1 (Kieft et al., 2020), VirNet v0.1 (Abdelkareem et al., 2018), Phigaro v2.2.6 (Starikova et al., 2020), VirSorter2 v2.0 (Guo et al., 2021), Seeker (Auslander et al., 2020) y MARVEL v0.2 (Amgarten et al., 2018). El desempeño de cada herramienta varía según el tipo de muestra, tecnología de secuenciación y método de ensamblaje.

Sin embargo, WtP recolecta las identificaciones positivas, las filtra utilizando un “threshold” para cada programa y así determinar un contig de fago positivo a partir de los resultados crudos de cada herramienta. Dichos programas están basados en diferentes estrategias de cálculo, bases de datos y dependencias, que se pueden resumir de la siguiente manera:

VirFinder: El primer método de “machine learning” basado en frecuencias k-mer para la identificación de contigs virales, sin utilizar similitud basada en genes para su búsqueda. Se basa en la observación empírica de que los virus y hospederos presentan diferencias perceptibles en sus patrones k-mers (Ren et al., 2017).

PPR-Meta: Un clasificador que permite simultáneamente identificar fragmentos de fagos y plásmidos, a partir de ensamblados metagenómicos. Utilizando “Deep learning”, entrenan y crean una arquitectura de red (llamada “Bi-path Convolutional Neural Network”), diseñada para mejorar el rendimiento de detección de fragmentos pequeños (Fang et al., 2019).

VirSorter: Herramienta diseñada para detectar señales virales en diferentes tipos de datos de secuenciación microbiana de manera dependiente e independiente de bases de datos, aprovechando modelos probabilísticos y extensos datos virómicos para maximizar la detección de nuevos virus (Roux et al., 2015).

DeepVirFinder: Basado en su predecesor (VirFinder), ahora (sin utilizar patrones predefinos como k-mers) diseña redes neuronales convolucionales para automáticamente aprender señales genómicas virales y simultáneamente construir un modelo predictivo basado en dichas firmas para predecir si una

secuencia es de un genoma viral (Ren et al., 2020).

Metaphinder: Método para identificar fragmentos genómicos ensamblados (por ejemplo, contigs) que provengan de fagos, en conjuntos de datos metagenómicos. Su estrategia se basa en la comparación con una base de datos de genomas completos de bacteriófagos, integrando hits a múltiples genomas para adaptarse a la estructura de mosaico de muchos bacteriófagos (Jurtz et al., 2016).

Sourmash: Instrumento para crear, comparar y manipular bocetos MinHash de datos genómicos. Estos bocetos permiten almacenar patrones de ADN y ARN de grandes colecciones de secuencias, y así buscar o compararlos para identificarlos en muestras o encontrar similares, utilizando el índice de similitud de Jaccard (Brown et al., 2016).

Vibrant: El primer método que utiliza un acercamiento híbrido de “machine learning” y similitud de proteínas que no depende de las características de las secuencias para una recuperación y anotación automática de virus, determinación de la calidad del genoma, completitud y caracterización de la función de la comunidad viral. Todo esto gracias a redes neuronales de patrones proteicos y una nueva métrica (v-score) que elude los límites tradicionales y maximiza la identificación de genomas virales líticos y provirus integrados (Kieft et al., 2020).

VirNet: Modelo de atención profunda (“machine learning”) para identificar lecturas virales de mezclas de secuencias virales y bacterianas, y purificar datos metagenómicos virales de contaminación bacteriana. Esto guiaría la

identificación de nuevos virus y potencialmente llevaría a cabo una caracterización funcional (Abdelkareem et al., 2018).

Phigaro: Paquete de Python con un algoritmo (PhigaroFinder) que define regiones de profagos putativos basado en un preprocesado de la información de entrada. Este preprocesado se lleva a cabo por dos programas externos, Prodigal que entrega una lista de genes con sus coordenadas, contenido GC y otras propiedades junto a las secuencias proteicas predichas. Después anota las secuencias de proteínas con HHMSCAN (Potter et al., 2018) utilizando modelos ocultos de Markov (HMMs) específicos para fagos de grupos ortólogos de virus procariontes (pVOGs) (Grazziotin et al., 2017). Un gen es considerado parecido a un fago si corresponde con uno de los PVOGs de los HMMs (Starikova et al., 2020).

VirSorter2: Herramienta para la identificación de virus de ADN y ARN que aprovecha los esfuerzos más recientes de secuenciación de los grupos virales subrepresentados para desarrollar clasificadores automáticos que mejoren la detección de los virus del orden *Caudovirales*, mientras que sigue identificando otros principales grupos virales de diversa longitud de genoma y hospederos (Guo et al., 2021).

Seeker: Método sin alineación que aprovecha los avances recientes en aprendizaje profundo para detectar fagos. A diferencia de otros métodos de aprendizaje de secuencias, Seeker emplea modelos largos de memoria a corto plazo que le permiten mantener una larga memoria de las secuencias, y por lo tanto puede identificar dependencias distantes dentro de una secuencia, para

distinguir fagos de bacterias (Auslander et al., 2020).

MARVEL: Herramienta para la predicción de secuencias de fagos de ADN de doble hebra utilizando un enfoque de “machine learning” y tres características genómicas simples (densidad de genes, cambios en la hebra y fracción de “hits” significativos en una base de datos de proteínas virales) extraídas desde los contigs para lograr un buen rendimiento al separar secuencias bacterianas con las de fagos (Amgarten et al., 2018).

5.7 Evaluación de la calidad de los genomas virales

Una vez que se completó la detección viral realizada por WtP para los 49 metagenomas a trabajar, se agruparon los contigs positivos para fagos y se les aplicó un filtro de calidad y completitud mediante la herramienta CheckV, la cual corresponde a un flujo de trabajo automatizado para evaluar la calidad de genomas virales que estén representados por un solo contig, estimando la completitud de fragmentos genómicos, identificando la contaminación del hospedero para secuencias que representen provirus integrados e identifica genomas cerrodos (Nayfach et al., 2021). Esto se logra siguiendo cuatro pasos principales. Primero, CheckV identifica y remueve regiones no virales en los provirus utilizando, un algoritmo que considera la anotación de genes (base de datos de HMM para genes marcadores) y el contenido GC. Luego, estima la completitud en base a una comparación con una gran base de datos de genomas virales completos derivados del GenBank del NCBI y muestras ambientales, utilizando AAI (“average amino acid identity”). Una vez que identifica los mejores “hits”, se entrega la completitud como una relación entre el largo del contig

ingresado y el largo de la referencia que hizo el “hit”, reportándose también un nivel de confianza según la fuerza del alineamiento. Con menor frecuencia puede ocurrir que el contig no tenga un match cercano con la base de datos de CheckV, para estos casos el programa estima la completitud en base a los HMMs virales identificados en el contig, donde CheckV entrega un rango estimado de completitud que representa el intervalo de confianza del 90% basado en la distribución de longitudes de genomas de referencias con los mismos HMMs virales. Finalmente, la predicción de genomas cerrados se basa en regiones flanqueadoras de hospederos, repetidos directos terminales o repetidos invertidos terminales, haciendo una referencia cruzada con las completitudes obtenidas previamente. Toda esta información se resume en un informe que asigna a cada secuencia uno de los cinco diferentes niveles de calidad en base a completitud (completo, alta calidad, media calidad, baja calidad y no determinado) (Nayfach et al., 2021). En este estudio se seleccionaron las secuencias virales de las primeras cuatro categorías, excluyendo los contigs que no se les logró determinar su calidad (no determinados).

5.8 Obtención de Unidades Taxonómicas Operativas de Virus

El siguiente paso para seguir con el minado viral corresponde a la generación de una población de contigs no redundantes. Para esto se recurrió a la herramienta bioinformática CD-HIT, un programa ampliamente utilizado para la clusterización de secuencias biológicas para reducir las redundancias y mejorar el desempeño de posteriores análisis de secuencias (Fu et al., 2012). Este programa presenta diferentes sub-programas según el tipo de trabajo que el usuario necesite y la

información de entrada que se le otorgue, siendo utilizado en este caso CD-HIT-EST, que clusteriza proteínas similares (ADNs) en clústeres que cumplen un umbral de similitud definido por el usuario. Para este caso, se estableció un 95% de identidad nucleotídica y un 80% de cobertura en relación con la secuencia más corta (Roux et al., 2017). Este umbral permite la formación de grupos de virus que presentan el mismo rango hasta el nivel taxonómico de especie, las llamadas Unidades Taxonómicas Operativas de Virus (vOTUs(Roux et al., 2019).

5.9 Obtención de Base de datos IMG/VR

El “Joint Genome Institute”, también conocido como JGI, reúne la experiencia y los recursos para desarrollar parte de los proyectos más grandes en mapeo genómico y secuenciación a nivel mundial. Uno de sus proyectos corresponde a la colección más grande de secuencias virales obtenidas desde metagenomas, la base de datos IMG/VR (Integrated Microbial Genomes/Virus), la cual recientemente publicó una nueva versión (cuarta versión) que cuenta con más de 15 millones de genomas virales y fragmentos genómicos (Camargo et al., 2022). Se filtró esta gran base de datos, seleccionándose únicamente los genomas virales no cultivados (UviGs) pertenecientes a sistemas acuáticos y termales dentro de un rango de temperatura de 25-90°C. Una vez obtenidas las secuencias de interés y su información, se clusterizan entre ellas para evitar redundancias, utilizando los mismos parámetros establecidos para los vOTUs, con la excepción de que, en esta ocasión, se clusterizan utilizando Mmseqs2 (Steinegger y Söding, 2017) para ahorrar tiempo de cómputo en consideración a la cantidad de datos por analizar.

5.10 Predicción de proteínas

Para futuros análisis taxonómicos, se predicen las proteínas de los vOTUs y los UviGs del IMG/VR, utilizando el programa Prodigal v2.6.3 (-p meta) (Hyatt et al., 2010). Comparaciones de proteínas distintivas pueden ser combinadas con métricas basadas en el contenido genético, orden genético y orientación (sintenia), además de otros aspectos de la organización del genoma y HMMs de las familias proteicas más conservadas para clasificar todos los virus procariontes a nivel de especie, género, subfamilia, familia y rangos taxonómicos superiores (Simmonds et al., 2023).

5.11 Asignación taxonómica de los vOTUs

La asignación taxonómica se realizó formando redes de genes compartidos basadas en clústeres de proteínas (PCs) entre genomas virales (en este caso, vOTUs y UViGs). Esta estrategia se ha demostrado que es, en gran medida, concordante con los taxones respaldados por el Comité Internacional de Taxonomía de Virus (Bolduc et al., 2017). Para formar esta red, se utiliza vConTACT v2.0, una aplicación basada en redes que utiliza perfiles de genes compartidos en genomas completos para taxonomía viral, integrando una clusterización jerárquica basada en distancias y puntajes de confianza para todas las predicciones taxonómicas (Jang et al., 2019).

Este programa requiere dos archivos de entrada para funcionar correctamente: Las proteínas crudas correspondientes a cada vOTU y UViG (este archivo corresponde a un documento de texto que contiene las secuencias aminoacídicas) que se obtuvieron previamente a través de Prodigal y un archivo

de asignación llamado "Gene to Genome" que se debe crear de manera previa utilizando el mismo programa vConTACT2 con las mismas proteínas mencionadas previamente. Este último archivo sirve para establecer el nexo entre los nombres de las proteínas y sus correspondientes genomas (vOTUs y UViGs).

Una vez obtenido esto, se ingresan los archivos y se configura vConTACT2, utilizando su configuración por defecto (--rel-mode 'Diamond' –db 'None' –pcs-mode MCL –vcs-mode ClusterONE), con la excepción de que se modificó la opción "db" que corresponde a la base de datos ocupada, eligiendo la alternativa de no utilizar una base de datos externa (--db 'None'). Esto último a raíz de que actualmente vConTACT2 utiliza la base de datos de RefSeq viral, que no se encuentra actualizada respecto a la nueva taxonomía liberada por el ICTV. Por esta razón, se utilizó en su lugar la nueva versión de la base de datos del IMG/VR (versión 4), que como se mencionó previamente, fue filtrada para sistemas termales y está descrita con la nueva taxonomía de ICTV (Camargo et al., 2022). Las proteínas predichas de esta base de datos se concatenan con las proteínas pertenecientes a los vOTUs, creándose el archivo de proteínas crudas de entrada para vConTACT2 y así formar clústeres entre vOTUs y los UViGs pertenecientes a la base de datos. vConTACT2 por su parte, ofrece distintas alternativas para afinar la asignación taxonómica. Por una parte, entrega un archivo con toda la información taxonómica de los genomas de referencia utilizados (genome_by_genome_overview), así como también toda la información respecto de la clusterización obtenida. Por otra parte, otro archivo crítico que entrega

(C1.NTW) contiene la información de los genomas virales introducidos en el programa y el peso de las conexiones que se realizaron entre ellos (que superan un umbral de significancia determinado por la probabilidad de que esos genomas conectados compartan una cierta cantidad de genes). Esta información se transfiere a programas de visualización de redes biológicas para que el usuario pueda crear una imagen adecuada para explicar sus resultados. En este caso, se utilizó el programa Cytoscape 3.9.1 (Kohl et al., 2011), una plataforma para visualizar redes complejas e integrarlas con cualquier tipo de datos de atributo. Inicialmente, con el primer resultado emitido por vConTACT2, se puede revisar el estado de clusterización de los vOTUs, existiendo 5: clusterizado, singleton, overlap, outlier y clusterizado/singleton. Aquí, los vOTUs “clusterizados” que pertenezcan a un mismo clúster que un genoma de referencia, muestra que existe una alta probabilidad de que el vOTU introducido por el usuario sea parte del mismo género que la referencia, por lo que se le asigna su misma taxonomía. Los “singleton” son secuencias que tienen muy pocas (o no presentan) similitudes con otros genomas virales, por lo que no se pueden asignar y la mayoría no aparecen en la red. “Overlaps” corresponde a genomas que presentan una superposición con genomas de múltiples clústeres virales. Esto suele ocurrir cuando hay un núcleo (core) génico compartido, o cuando una gran parte de su genoma presenta una región conservada entre varios de ellos, por lo que también son asignables según el VCs al cual se encuentren relacionados. Los “Outliers” presentan genes compartidos con otros genomas que pertenezcan a un VC, sin embargo, ClusterONE no tuvo la confianza suficiente para ubicarlo dentro de

dicho VC. En este caso la sospecha es que efectivamente están relacionados al VC al cual están conectados (visible en la red), pero no al nivel de género, probablemente a nivel de familia o sub-familia, por lo que hasta este rango taxonómico se pueden asignar al rescatarlos en la red de genes compartidos. Finalmente, están los Clusterizados/Singleton, esta categoría es un tanto particular, ya que son genomas que ClusterONE ubicó en un VC, sin embargo, cuando se aplicó un umbral basado en la distancia con genomas de referencia, vConTACT2 decide que no están en el mismo género y los mueve a un sub-clúster, pero que cuando este genoma viral no tiene ningún otro que se mueva a este nuevo sub-clúster, queda solo (por esto se le determina también como singleton). Básicamente es un singleton que, sí está clusterizado, pero su clúster fue separado, por lo que no se asignan taxonómicamente.

Finalmente, para la creación de la red, se pueden utilizar distintos algoritmos que determinan la distribución espacial de acuerdo con una función específica, siendo utilizado en este caso un diseño llamado “Edge-Weighted Spring- Embedded Layout”, basado en un paradigma basado por la fuerza, donde los genomas virales (nodos) son tratados como objetos físicos que se repelen entre ellos, como si fuesen electrones. Mientras, las conexiones entre los nodos (“edges”) se tratan como resortes metálicos atados a los nodos, repeliendo o atrayendo sus extremos de acuerdo con una función de fuerza. Este algoritmo ubica los nodos en una posición tal que se minimice la suma de las fuerzas en la red (Kamada y Kawai., 1988).

5.12 Análisis de Abundancias de vOTUs en los metagenomas

De manera paralela, se realizaron análisis de abundancias a partir del set de vOTUs obtenidos tras la clusterización realizada por CD-HIT. Para esto, se recurrió al programa bowtie2, una herramienta que permite alinear secuencias utilizando índices para hacer más rápido y eficiente el proceso, mientras que permite que el alineamiento contenga brechas, separando el proceso en dos partes: Un alineamiento inicial sin brechas que permite buscar la posición de cada contig dentro de una lectura limpia y una extensión con brechas que permite el alineamiento completo del contig, considerando dichas brechas (Langmead y Salzberg, 2012). Se procedió a mapear el set de vOTUs contra las lecturas limpias (forward y reverse) de cada metagenoma en este estudio, y posteriormente se utilizó la herramienta SAMtools (Danecek et al., 2021) para procesar y analizar los resultados del alineamiento.

A partir de estos archivos, todo el trabajo posterior y procesamiento se realizó utilizando el lenguaje de programación R (4.2.2), a través de la plataforma RStudio. Brevemente, el flujo de trabajo fue el siguiente:

Primero se realiza un filtro, para dejar solamente a los vOTUs que tengan una cobertura mayor al 75% dentro de las lecturas y con estos vOTUs se construye una tabla de conteos. Posteriormente, se normaliza la tabla de conteos por TPM (transcrito por millón) utilizando el paquete de R edgeR (Robinson et al., 2010), para reducir posibles parcialidades debido a los distintos tamaños de cada metagenoma. Una vez que se tiene la tabla de conteos normalizada, se utiliza el paquete de R Phyloseq (McMurdie y Holmes, 2013) para lograr una

representación y análisis interactivo, además de reproducible, respecto de las abundancias de los vOTUs por metagenomas y de qué manera están distribuidos. De manera paralela, se alineó el set de vOTUs de El Tatio con las lecturas limpias de cada metatranscriptoma de El Tatio (13) con bowtie2 (Langmead y Salzberg, 2012), y procesó el resultado con SAMtools (Danecek et al., 2021). Luego, se aplica el mismo procedimiento descrito previamente con el lenguaje de programación R (4.2.2) y RStudio para llegar a la tabla de conteos normalizada por TPM y objeto phyloseq correspondiente a los vOTUs provenientes de El Tatio en sus metatranscriptomas. La única excepción consiste en que, para el procesamiento de los datos del reclutamiento, no se aplicó el filtró inicial de cobertura de los vOTUs en las lecturas limpias (>75%), que sí se aplica en el caso de los vOTUs en sus metagenomas.

5.13 Predicción Virus-Hospedero

Los hospederos microbianos para las comunidades virales en El Tatio se predijeron utilizando una gran variedad de métodos bioinformáticos que incluyen “matchs” virales con espaciadores CRISPR en el hospedero, profagos integrados en el genoma del huésped, genes de ARNt en el hospedero y patrones k-mers en el hospedero. Todo esto lo aúna una herramienta llamada VirMatcher (Bolduc y Zayed, 2020), disponible mediante la plataforma online KBase, donde se pueden ingresar genomas correspondientes a hospederos (MAGs) y genomas virales (vOTUs). Luego, VirMatcher aplica las 4 estrategias mencionadas para la predicción de relaciones virus-hospederos y les asigna un “score” de confianza según el rendimiento de cada plataforma en conjunto.

Los 4 acercamientos que utiliza el programa están basados en:

minCED: Herramienta basada en la búsqueda de repeticiones cortas exactas de largo *k* que estén separadas por una distancia similar para luego extender estas coincidencias exactas de *k*-meros a la longitud de repetición real. Una vez que se encuentran repeticiones reales, son filtradas para eliminar aquellas que no cumplan con los requisitos específicos de CRISPR (Bland et al., 2007).

Finalmente, se aplica un BLASTn para evaluar coincidencias entre los espaciadores CRISPR y las poblaciones virales de El Tatio.

Blast de profagos: Se realiza un BLASTn (-task megablast) de los virus obtenidos en El Tatio contra los MAGs correspondientes a los mismos sistemas termales utilizados para el minado viral. De esta manera, los genomas microbianos con regiones de su genoma con un valor mayor o igual a 2500 pares de bases que coincidan en un 90% de ID con algún genoma viral de El Tatio se conservan (Roux et al., 2016). Estos van a ser filtrados por cobertura del contig viral (requiriendo al menos un 30% de cobertura viral) y cobertura del MAG hospedero (requiriendo que al menos 30% del MAG esté afuera de la región de alineación del prófago para evitar fragmentos virales mal clasificados en los genomas microbianos). Por último, se asocia un puntaje a los contigs virales según su cobertura (Gregory et al., 2020).

tRNAscan-SE: Programa que identifica 99-100% de genes de ARN de transferencia (tRNA) en secuencias de ADN o ARN en tres fases. En la primera etapa, analiza las secuencias con tRNAscan (Fichant y Burks., 1991) y EufindtRNA (Pavesi et al., 1994) y añade los resultados en un listado de tRNAs

candidatos. En la segunda etapa, tRNAscan-SE extrae las subsecuencias de los candidatos y transfiere estos segmentos a un programa de búsqueda de modelos de covarianza (*covels*) (Eddy y Durbin., 1994), agregando 7 nucleótidos flanqueantes en ambas partes de los tRNAs candidatos en caso de que el tRNA haya sido truncado por la predicción inicial. Finalmente, tRNAscan-SE toma los tRNAs que han sido confirmados con *covels,* que registran un puntaje mayor a 20 bits, los recorta y aplica el programa de alineación de la estructura global del modelo de covarianza, *coves* (Eddy y Durbin., 1994) para obtener una predicción de estructura secundaria. El isotipo del tRNA se predice al identificar el anticodón dentro del resultado de estructura secundaria de *coves*, mientras que los intrones se identifican como series de cinco o más nucleótidos sin consenso consecutivos dentro del bucle de anticodón (Lowe y Eddy., 1997). Posterior a esto, se realiza un BLASTn entre los genes de tRNA virales y microbianos. También se buscan genes de tRNA con BLASTn contra las secuencias de tRNA del conjunto de datos virómicos de la Tierra (Paez-Espino et al., 2016), eliminando tRNAs promiscuos y asignando un puntaje a las coincidencias entre los virus y sus hospederos de acuerdo con su exactitud.

WIsH: Herramienta que predice los hospederos procariotas de contigs virales mediante el entrenamiento de un modelo homogéneo de Markov de orden 8 para cada posible genoma hospedero, luego calcula la probabilidad de un contig bajo cada uno de los modelos de Markov entrenados y predice *de novo* el hospedero cuyo modelo produce la mayor probabilidad (Galiez et al., 2017). Este programa se utiliza después de enmascarar secuencias de tRNA en los

genomas virales para mejorar el rendimiento (Gregory et al., 2020).

5.14 Análisis estadístico de diversidad

Utilizando el nuevo set de vOTUs (tras su filtración por cobertura en las lecturas) y su correspondiente tabla de conteos normalizada, se crea un nuevo objeto phyloseq, sólo que, en esta ocasión, también se añade al objeto la tabla de metadata correspondiente a cada metagenoma (Tabla suplementaria 1).

Primero se realizaron análisis de diversidad alfa, para ver cómo varía la diversidad de los vOTUs en cada muestra (metagenoma), determinando cuáles son los factores que mejor explican dichos cambios. Para esto, se utilizaron los índices de Shannon (bajo una función exponencial), Simpson, Chao1 (también conocido como Riqueza observada) y Equidad de Pielou. Se calculan los índices para cada muestra utilizando phyloseq y se tabulan junto a la metadata correspondiente. Luego, se aplicó una prueba de Shapiro-Wilk para medir la normalidad de los índices de interés (si p-value > 0,05, se asume normalidad). Finalmente, para las variables continuas (Temperatura, pH y Altitud) se utiliza un modelo lineal generalizado (GLM) para el cual se estima la significancia del modelo completo mediante ANOVA, mientras que para las variables categóricas (Origen de la muestra, hábitat) se usa un test no paramétrico de Kruskal-Wallis, visualizándose los resultados con el paquete de R llamado visreg 2.7.0.3 (Breheny y Burchett., 2013).

Posteriormente se realizaron los análisis de diversidad beta, para observar la partición de la diversidad biológica entre sitios o a lo largo de un gradiente. En este caso, se extrajeron las variables ambientales del objeto phyloseq

previamente mencionado y a las variables numéricas se les realiza un centrado y escalado, para finalmente crear un nuevo objeto phyloseq que tenga dichas variables con esta nueva modificación. A este nuevo objeto phyloseq se le procede a realizar un “forward selection”, un tipo de regresión progresiva que comienza con un modelo vacío y va agregando las variables una a una y va calculando (prueba F de Fisher) la contribución de cada una al modelo, de manera que finalmente permite determinar qué variables son las que más aportan y que tan significativas son. Los parámetros que se rescataron de esta selección son utilizados para realizar un análisis de redundancia (RDA) a través del paquete de R ampvis2 (Andersen et al., 2018), y así resumir la variación en el set de muestras y su respuesta frente a los parámetros más determinantes de su composición/distribución.

De manera paralela, se identificó la fuente de varianza considerando las variables no numéricas (categóricas), considerando además la ubicación (coordenadas UTM, Universal Transverse Mercator) y la fuente del ADN (tapete, sedimento o agua), mediante un Análisis de varianza multivariado permutacional (PERMANOVA) (Anderson, 2000) con la función adonis2 (que utiliza matrices de distancia, agregando los términos de manera no secuencial) del paquete de R vegan (Oksanen et al., 2020). Reajustando el RDA con las nuevas variables, según significancia y porcentaje de la varianza.

5.15 Análisis de diversidad basado en distancias

Finalmente, se emplearon análisis de diversidad (beta) de coordenadas para el índice de disimilitud de Bray-Curtis mediante Análisis de coordenadas principales

(PCoA) y Escalamiento multidimensional no métrico (NMDS). Estas ordenaciones sin restricciones basadas en la distancia buscan representar la (di)similitud entre objetos (muestras, sitios) en base a los valores de múltiples variables (columnas) asociadas a ellos, de manera que los objetos similares son representados cerca uno del otro y lo disimilares se encuentras alejados entre sí. Estos análisis exploratorios multivariados son útiles para revelar patrones en grandes grupos de datos (Ramette, 2007).

Para ambos casos, se utiliza el paquete de R ampvis2 (Andersen et al., 2018), un envoltorio que trabaja alrededor del paquete de R vegan (Oksanen et al., 2020), aplicando la medida de distancia “bray” apoyada por la función vegdist de vegan. No se realizan transformaciones a las abundancias antes de la ordenación, sólo se filtran las especies menos abundantes (menor al 0,1%). El valor final se representa mediante un objeto ggplot2 (Wickham, 2014) para generar gráficos de ordenación adecuados para el análisis y comparación de comunidad microbianas (Andersen et al., 2018).

6 Resultados

6.1 Composición y diversidad de las comunidades virales de ADN presentes en tapetes microbianos de termas en El Tatio.

6.1.1 Identificación de secuencias virales desde el catálogo de contigs de los metagenomas de tapete microbiano de termas de El Tatio.

Los géiseres de El Tatio están localizados en los montes andinos del norte de Chile, en la región de Antofagasta (22°19′53″S 68°0′37″O), a unos 4200 metros sobre el nivel del mar (msnm) (Figura 1a). Parámetros fisicoquímicos de 13 sitios distintos de muestreo en las termas de El Tatio fueron medidos en enero del año 2020 (Tabla suplementaria 1, Figura 1b y c). La temperatura de la mayoría de las muestras está alrededor de los 55°C y un pH que oscila entre 6-9 (Figura 1b y c). Los metagenomas analizados corresponden a muestras de ADN extraído directamente del tapete microbiano de los 13 sitios mencionados en termas de El Tatio, las cuales fueron secuenciadas mediante Illumina Novaseq 6000 (Roy J. Caver Biotechnology Center, Univesity of Illinois, USA). Los archivos correspondientes a las lecturas se filtraron por calidad acorde a Guajardo-Leiva et al., (2018). El trabajo realizado en este proyecto inicia a partir del ensamble de las lecturas correspondientes a la secuenciación de las muestras de ADN de enero del 2020, donde se partió con la identificación de los contigs de procedencia viral.

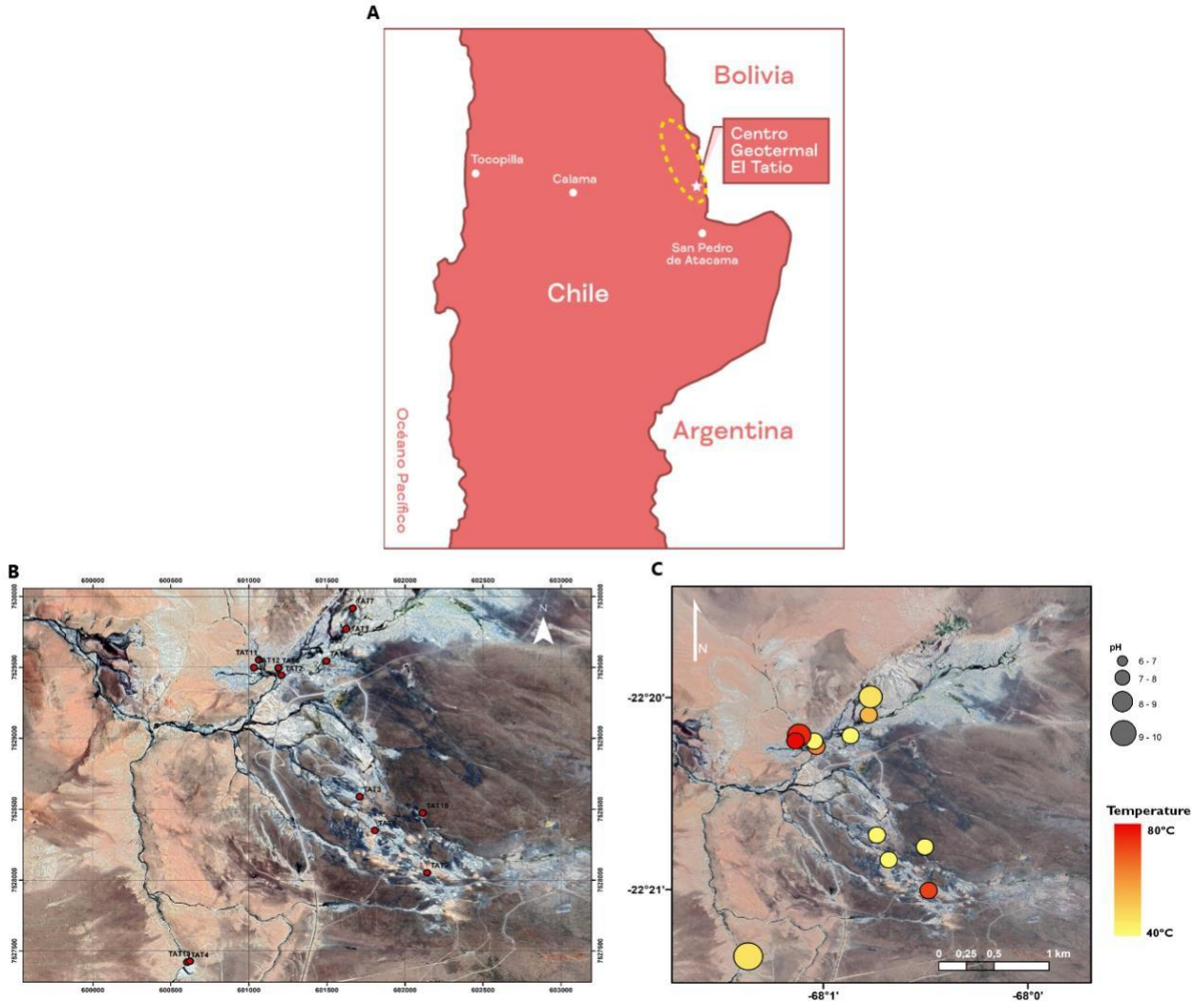

Figura 1.- Mapa de los géiseres de El Tatio, San Pedro de Atacama, Región de Antofagasta, Chile.

(A) Ubicación geográfica a nivel regional del campo geotermal de El Tatio. Localización de los 13 puntos de muestreo (termas) en el interior del campo geotermal de El Tatio, denominándose en el mapa como TAT1-13 (detalles sobre coordenadas se encuentran en la Tabla suplementaria 1).

Temperatura y pH de los 13 puntos de muestreo en El Tatio, representado cada sitio en función de un gradiente de colores y círculos de tamaño proporcional, respectivamente.

Para la identificación de secuencias virales, primero se realizó un filtro de los contigs por tamaño (> 10000 pb), y luego se procedió a la detección viral con WtP (Marquet et al., 2022) correspondiente a los 13 metagenomas pertenecientes a El Tatio. Se logró identificar un total de 2061 contigs virales que cumplieron con el umbral asignado por WtP para detección viral. Sin embargo, solamente 705 contigs pasaron el filtro de calidad realizado por CheckV (Nayfach et al., 2021), es decir, entraron en las categorías de alta, media y baja calidad (quedando fuera los contigs de la categoría “no determinado”). Estos 705 contigs se clusterizaron (95% de identidad, 80% de cobertura) para crear el set no redundante de vOTUs pertenecientes a El Tatio, obteniéndose finalmente 527 vOTUs.

6.1.2 Identidad taxonómica de vOTUs de tapetes microbianos de las termas en El Tatio.

La asignación taxonómica de los vOTUs provenientes de las 13 termas en El Tatio se realizó a través de redes de genes compartidos basadas en clústeres de proteínas compartidos mediante la herramienta vConTACT2 (Jang et al., 2019). Inicialmente se utilizó la base de datos RefSeq viral, versión 201, del NCBI. Sin embargo, debido a la baja asignación que se puede realizar utilizando esta base de datos, y a la actualización y cambios realizados por el ICTV a la taxonomía de los virus, es que se optó por utilizar la base de datos del IMG/VR versión 4 (Camargo et al., 2023) de sistemas termales. Esta nueva base de datos es más extensa y permite manejar de mejor manera los datos metagenómicos, y esta actualizada con la nueva taxonomía viral.

Además, debido a la baja cantidad de vOTUs totales en el set proveniente de los 13 metagenomas de El Tatio y la naturaleza de vConTACT2 para realizar asignaciones taxonómicas (formación de clústeres virales según similitud entre sus proteínas), se optó por realizar el análisis de redes de genes compartidos de vConTACT2 utilizando un nuevo set global de vOTUs termales que en total representan 49 metagenomas (Los 13 metagenomas de El Tatio mencionados previamente, 4 metagenomas más de El Tatio disponibles en nuestro laboratorio y 32 de otras ubicaciones geográficas alrededor del mundo) (Figura 2). De esta manera, se facilita la formación de clústeres virales (VC) entre las secuencias virales de El Tatio y las referencias del IMG/VR, al permitir la formación de más clústeres y de mayor tamaño, y por tanto más robustos ya que, al haber más secuencias externas a El Tatio, pueden actuar como una especie de “puente” génico para unir sub-clústeres que puedan ser de El Tatio e IMG/VR respectivamente, y formar un VC más grande que pueda ser asignado taxonómicamente (fenómeno también llamado superposición de VCs). También permite una formación de VCs más grandes al generar un “core” génico más grande, atrayendo outliers o más nodos a que sean parte de dicho VC.

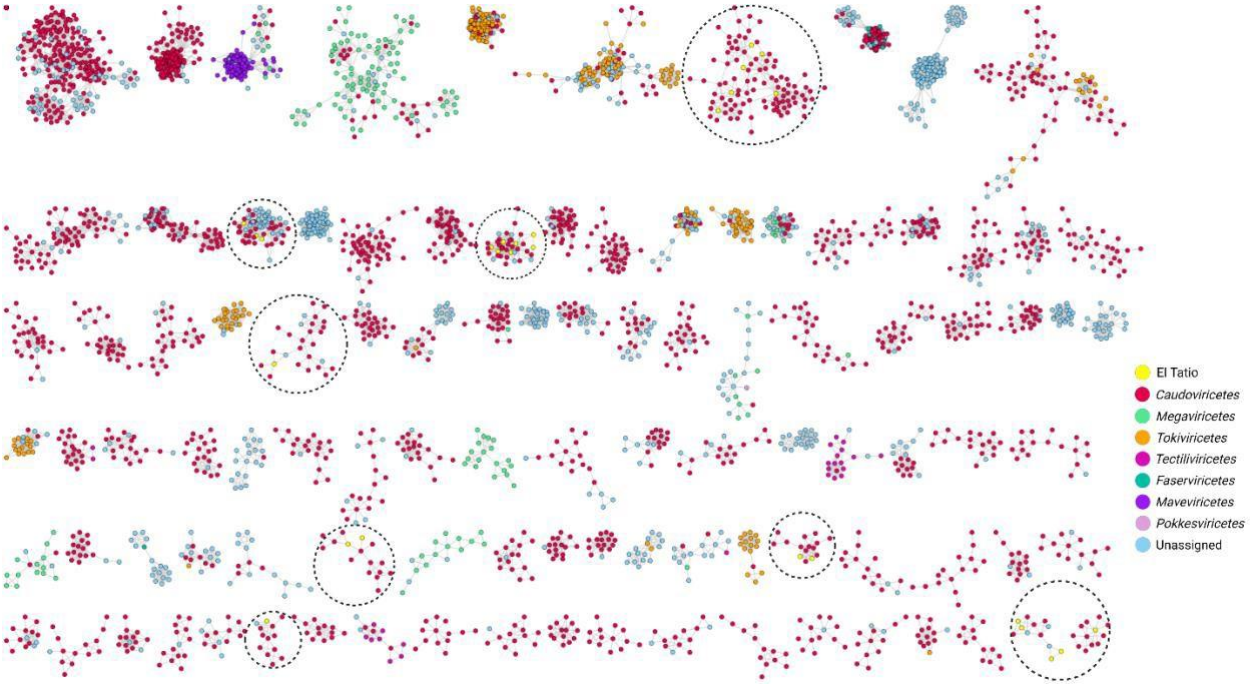

Figura 2.- Red de genes compartidos entre el set global de vOTUs (49 metagenomas) y la base de datos del IMG/VR v4 de sistemas termales.
Se construyó una red biológica de genes compartidos con Cytoscape 3.9.1 (Kohl et al., 2011), utilizando 67588 genomas de virus recuperados de la base de datos IMG/VR v4 de sistemas termales y 3559 vOTUs correspondientes al set global de este estudio. En esta red, los nodos (círculos, representando las proteínas de un genoma viral/vOTU/UViG) se colorearon según sus asignaciones taxonómicas a nivel de clase viral, distinguiendo particularmente los vOTUs de El Tatio. Su distribución espacial se debe al algoritmo “Edge-Weighted Spring Embedded Layout” (Kamada y Kawai., 1988). Los nodos se unen y forman agrupaciones (VCs) mediante líneas o conexiones (“edges” o “vértices”) obtenidas con vConTACT2 0.9.19 (Jang et al., 2019) y representan la conexión que vConTACT2 les asigna en base a una función de probabilidad según la cantidad de proteínas compartidas. Esto permite observar a qué grupos taxonómicos se asocian (o distancian) las secuencias, remarcándose en este caso, los VCs donde hay vOTUs de El Tatio presentes mediante circunferencias negras y así observar cómo se agrupan con las bases de datos que más parecido tienen con ellos y la distribución que dichas clases virales presentan.

Por lo general, no todos los vOTUs, UViGs o genomas virales logran incorporarse en la red (por ejemplo, todos los singletons quedan fuera de la red), razón por la cuál en la figura 2, al colorearse los vOTUs pertenecientes a El Tatio, representan un número mucho menor de lo que realmente vConTACT2 fue capaz de asignar. Además, esta red fue recortada para lograr la mejor visualización posible. La red poco a poco va perdiendo las grandes agrupaciones de nodos con taxonomía asociada, terminando por ser muchos pequeños grupos sin taxonomía asociados entre sí, resultando en muchos VCs de 2 o 3 nodos, que, si bien no se logra hacerles asignaciones taxonómicas, su distribución espacial y agrupaciones puntuales permiten estimar qué secuencias son más parecidas entre sí, pero aún desconocidas taxonómicamente. Muchos de estos nodos pequeños fueron eliminados de esta red para su mejor visualización en esta figura. En la Figura suplementaria 1 aparece la red completa.

En la figura 2, se muestran, encerrado en círculos negros, a los clústeres donde están presentes vOTUs provenientes de El Tatio, evidenciándose en cada caso la directa asociación y agrupamiento a genomas virales correspondientes a la nueva Clase de los *Caudoviricetes*. A esta clase ahora pertenecen los bacteriófagos y virus de arqueas con cola y cabeza, previamente designados cómo orden *Caudovirales* y sus tres familias: *Myoviridae*, *Podoviridae* y *Siphoviridae*. Sin embargo, la clasificación dentro de la clase de los *Caudoviricetes* hacia niveles más bajos de taxonomía no ha sido completamente formalizada, y se encuentra en un estado de descubrimiento y refinamiento (Evseev et al., 2023). Por esta razón, en la red se optó por usar el rango

taxonómico de Clase, ya que ninguno de estos vOTUs de El Tatio observados en la red (Figura 2) que pertenecen a la clase *Caudoviricetes* tuvieron afiliaciones directas (edge) a nuevos órdenes o familias virales.

De todas maneras, como se mencionó previamente en la metodología, Cytoscape utiliza un algoritmo particular para establecer la distribución de esta red, el modelo edge-weighted spring embedded. Este modelo ubica a los genomas o fragmentos que comparten más clústeres de proteínas más cerca entre sí, por lo que la distribución espacial toma gran relevancia, a diferencia de cuando se utilizaban las anteriores bases de datos y taxonomía viral, donde los grandes grupos virales como el abolido orden *Caudovirales* se asociaban todos en una nube gigante de nodos compuestas por gran cantidad de VCs pertenecientes a sus 3 familias ya mencionadas. Sin embargo, con los recientes cambios de taxonomía y la gran actualización de la base de datos del IMG/VR, los *Caudoviricetes* se encuentran mucho más dispersos ahora en la red, permitiendo que los vOTUs provenientes de El Tatio pudiesen distribuirse de manera mucho más heterogénea mientras que conservan su asignación como *Caudoviricetes*. Esto permite a su vez, observar grupos de vOTUs de El Tatio cercanos a otras clases virales en ocasiones, como ocurre con los *Tokiviricetes (*Figura 2*)*.

Considerando únicamente los vOTUs que se obtuvieron del set proveniente de los 13 metagenomas de El Tatio (excluyendo los 4 metagenomas extras de El Tatio que se agregaron posteriormente para el análisis taxonómico y global), 181 fueron asignados taxonómicamente al menos hasta el nivel de clase,

predominando los *Caudoviricetes* (99.5%), seguido de los *Tokiviricetes* (0.5%). La clase de los *Caudoviricetes* se asignaron en una pequeña proporción al orden de los *Thumleimavirales* (1.7%), y en la misma cantidad a la familia *Hafunaviridae*, mientras que la mayoría restante no pudo ser asignado. Por otra parte, el 0.5% de lo asignado a la clase *Tokiviricetes*, a su vez pudo asignarse en su completitud al orden *Ligamenvirales* y la familia *Rudiviridae*. Toda esta asignación correspondería a un 34.3% del total de vOTUs encontrados en El Tatio. De todas maneras, esta información no considera la abundancia que cada uno de estos virus asignados presenta en los 13 metagenomas de El Tatio. Para lograr esto, se calculó la abundancia relativa de cada uno de los vOTUs obtenidos en los 13 metagenomas (tanto asignados como no asignados taxonómicamente) y se representó la taxonomía de los virus de El Tatio según sus abundancias para dimensionar la cantidad de secuencias virales conocidas y desconocidas, que se encuentran presentes en cada metagenoma (Figura 3 y 4).

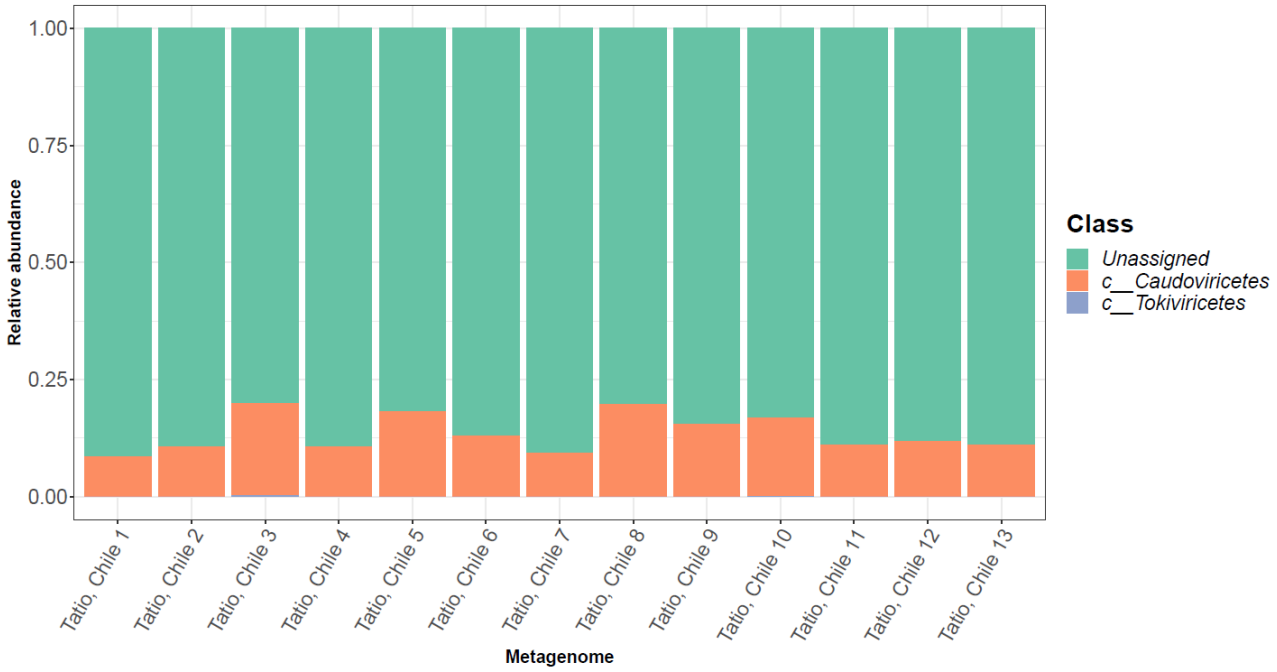

Figura 3.- Abundancia relativa por afiliación taxonómica de los vOTUs de El Tatio (13 metagenomas) a nivel de Clase. Gráfico de composición de barras apiladas que representa la abundancia relativa de los vOTUs correspondientes a los 13 metagenomas de El Tatio que fueron asignados taxonómicamente a nivel de clase. La asignación taxonómica se realizó mediante redes de genes compartidos utilizando los vOTUs de los 49 metagenomas más la base de datos IMG/VR versión 4 (2022), para ampliar en lo posible el número de asignaciones a rescatar desde vConTACT2.

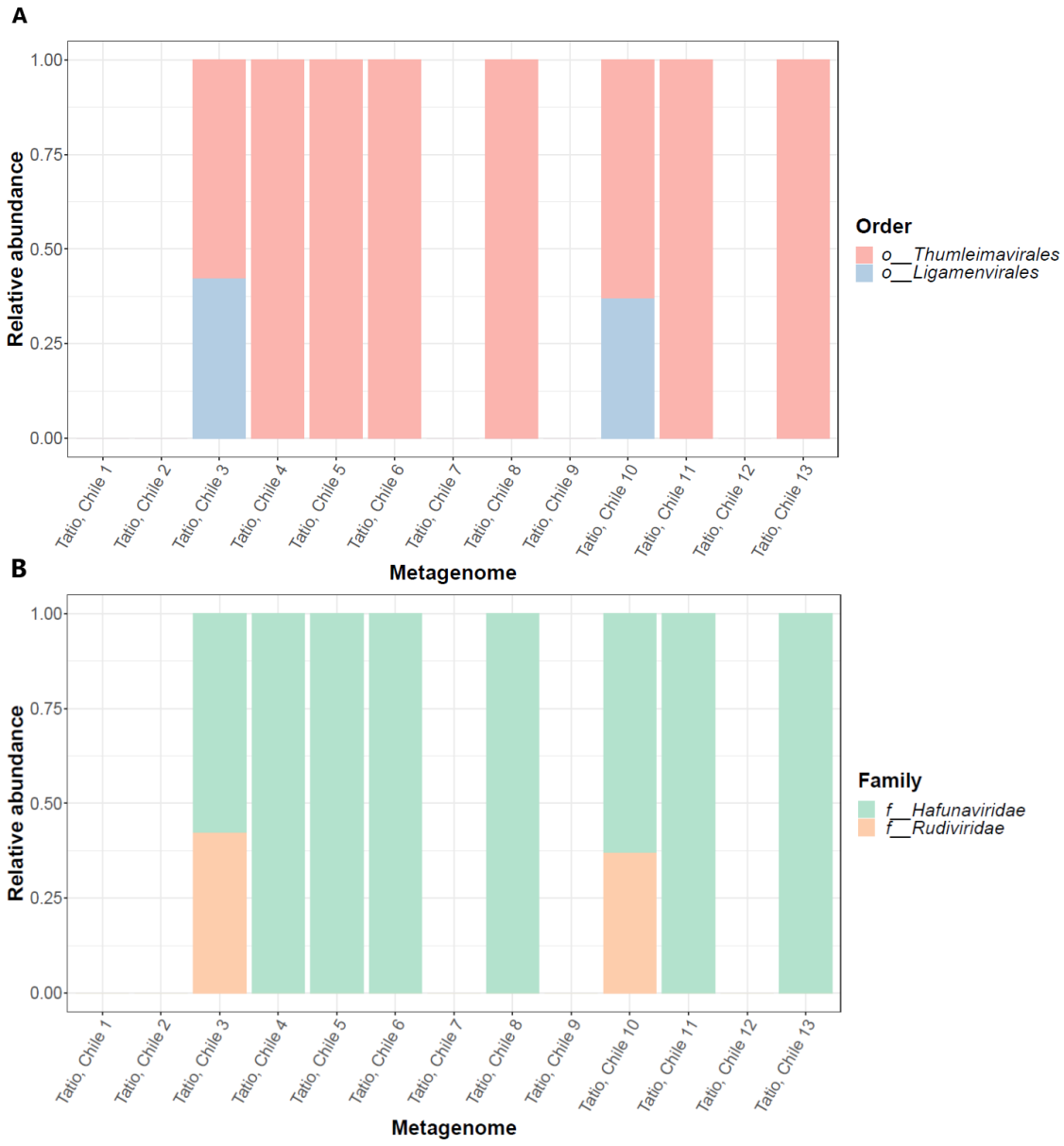

Figura 4.- Abundancia relativa por afiliación taxonómica de los vOTUs de El Tatio (13 metagenomas) a nivel de Orden y Familia.

Gráficos de composición de barras apiladas que representan la abundancia relativa de los vOTUs correspondientes a los 13 metagenomas de El Tatio.

(A) Afiliación taxonómica a nivel de Orden viral. (B) Afiliación taxonómica a nivel de Familias virales. Para ambos casos, se procedió con la asignación con vConTACT2 como se describió en la Figura 3. Las muestras para las que no se logró realizar ninguna asignación taxonómica no fueron coloreadas.

A partir del set de vOTUs de El Tatio (originarios de los 13 metagenomas) y teniendo en cuenta cómo y cuánto están distribuidos en estos metagenomas, se observó que en ningún caso se logró superar un 25% de asignación taxonómica del total en cada muestra (Figura 3). De manera más clara se puede evidenciar la presencia de la clase *Tokiviricetes* cuando se avanzó de rango taxonómico a nivel de Orden, donde para las mismas muestras (metagenomas 3 y 10), es visible la presencia del orden *Ligamenvirales* (de la clase *Tokiviricetes*), mientras que para el resto de las muestras en El Tatio (incluyendo la 3 y 10) continúa la predominancia de la clase *Caudoviricetes*, correspondiendo ahora al nuevo orden viral *Thumleimavirales* (Figura 4a). Finalmente, en el rango taxonómico de familia, al observar los metagenomas que tuvieron asignaciones taxonómicas con los que no lograron ser asignados, se encontró que al igual que lo encontrado a nivel de orden viral, los metagenomas 1, 2, 7, 9 y 12, no tienen una taxonomía asociada (Figura 4b), siendo en estos casos posible solamente asociar sus virus a un rango taxonómico mayor, la clase *Caudoviricetes*. No obstante, para los metagenomas que, si se pudieron asignar a nivel de familia, hay predominancia de la familia *Hafunaviridae* proveniente del orden *Thumleimavirales*, mientras que también aparece la familia *Rudiviridae*, proveniente del orden *Ligamenvirales* y que presentan similar abundancia relativa al ser los mismos virus los representados (Figura 4a y 4b).

6.1.3 Diversidad y abundancia de las secuencias virales provenientes de distintas fuentes termales en El Tatio.

En la ecología microbiana, los análisis de diversidad y abundancia de datos de secuenciación son un enfoque común para evaluar las diferencias entre entornos, ver patrones que se puedan traducir en la función del ecosistema o comprender cómo y por qué la diversidad cambia en el espacio y el tiempo.

Se estudió la diversidad alfa en las termas de El Tatio, donde se encontró que los datos seguían una distribución normal para los índices utilizados (Shannon, Pielou, Chao1 y Simpson), acorde a la prueba de Shapiro-Wilks aplicada (p-value fue mayor a 0.05; Figura suplementaria 2). Esto nos permitió proceder directamente a la aplicación de los GLM y ANOVA en las distintas variables ecológicas disponibles (temperatura, pH, latitud y longitud). Se apreciaron diferencias significativas en la diversidad alfa para la equidad de Pielou y equidad de Shannon (Figura 5). También se indagó en la discretización de los datos mediante gráficas de siluetas para calcular los números óptimos de clústeres para agrupar y k-means (método de agrupamiento) para las variables continuas temperatura y pH, pero tampoco se apreciaron diferencias significativas en la diversidad alfa de los grupos (se agrupó hasta en 4 grupos distintos).

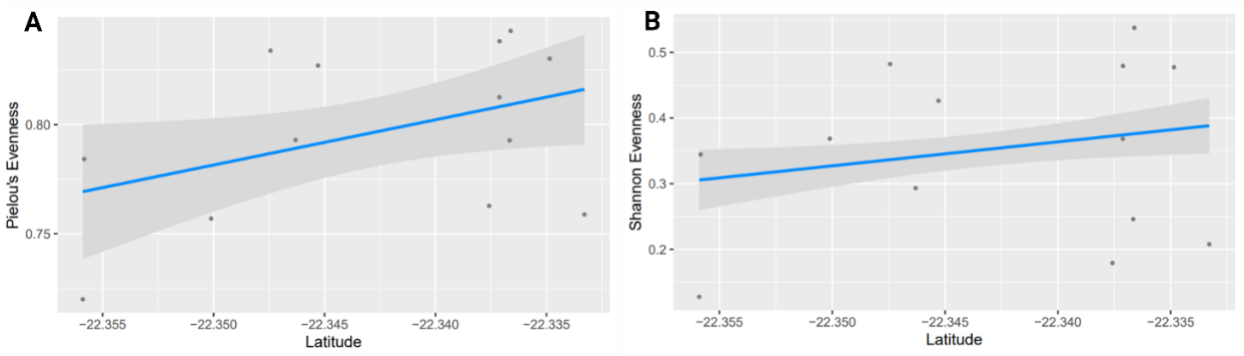

Figura 5.- Modelo Lineal Generalizado (GLM) simple en un gradiente de latitud, para los 13 metagenomas de El Tatio, utilizando los índices de diversidad alfa Shannon (Equidad) y Equidad de Pielou.

(A) GLM aplicado al gradiente latitudinal, de acuerdo con la equidad de Pielou como índice de diversidad alfa, obteniéndose un p-value de 0.05099.

GLM aplicado al gradiente latitudinal utilizando el índice de Shannon, aplicándose una función exponencial para representar el cambio de diversidad alfa de manera creciente. Particularmente se utilizó la equidad de Shannon, obteniéndose un p-value de 0.03501.

Los resultados de alfa diversidad indican que a medida que se recorre latitudinalmente el campo geotermal de El Tatio y sus respectivas fuentes termales, el total de especies presentes van evidenciando un aumento en la similitud de su distribución. Es decir, se muestra un aumento en la equidad de las distintas especies virales de manera significativa, a medida que se avanza desde el sur hacia el norte a lo largo de su extensión (Figura 5).

Por otro lado, se realizó un análisis de beta diversidad. Primero se realizó una regresión paso a paso, particularmente un *forward selection* utilizando distancias de Bray-Curtis junto a los parámetros que se podían centrar y escalar, es decir, pH, temperatura, altitud, latitud y longitud. Sin embargo, el resultado del modelo global no resultó ser significativo, por lo que no se siguió adelante, ya que ningún parámetro sería significativo, como para graficar en un análisis de redundancia o un análisis de correlación canónica (utilizándolos como variables constreñidas). De manera complementaria, se recurrió a la función adonis2 del paquete de R vegan (Oksanen et al., 2020) para hacer un análisis permutacional de varianza (PERMANOVA), donde también en base a distancias de Bray-Curtis y los parámetros ambientales previamente mencionados, se testeó la significancia de los factores ambientales en la varianza de los datos, siendo estos añadidos independientemente y de manera no secuencial al modelo. Se obtuvo que la longitud explicaba el mayor porcentaje de la varianza (19.8%), con un p-value de 0.027. Seguido a este factor se encontró a la temperatura con una varianza del 17.8% (p-value = 0.047). El resto de las variables ambientales testadas no mostraron diferencias significativas ni mayor porcentaje de la varianza.

Además, se realizaron análisis de diversidad beta de coordenadas utilizando la ocurrencia y abundancia de los vOTUs por metagenoma para calcular los índices de disimilitud de Bray-Curtis. Finalmente se optó por realizar un Análisis de Coordenadas Principales (PCoA), utilizando la transformación “sqrt” (Square Root Transformation) para transformar el índice de disimilitud no métrico a uno métrico (Figura 6). A nivel de vOTU, las muestras que presentan un rango de temperatura parecido (~55°C), fueron más similares entre sí, demostrando mayor cercanía, independiente de su pH. Cuando la temperatura disminuye (a 50°C), empieza a disminuir la similitud con el primer grupo y se alejan en dirección del punto de menor temperatura de la ordenación (45°C), el cual está evidentemente distanciado de las grandes agrupaciones que presentan temperaturas y cercanías similares (~55°C), denotando una diferencia en sus vOTUs en comparación al resto de los metagenomas. La Terma 9 fue la más disimilar al resto, a pesar de presentar un pH y temperatura similar cercana a los 55°C (54.3°C) y un pH relativamente neutro.

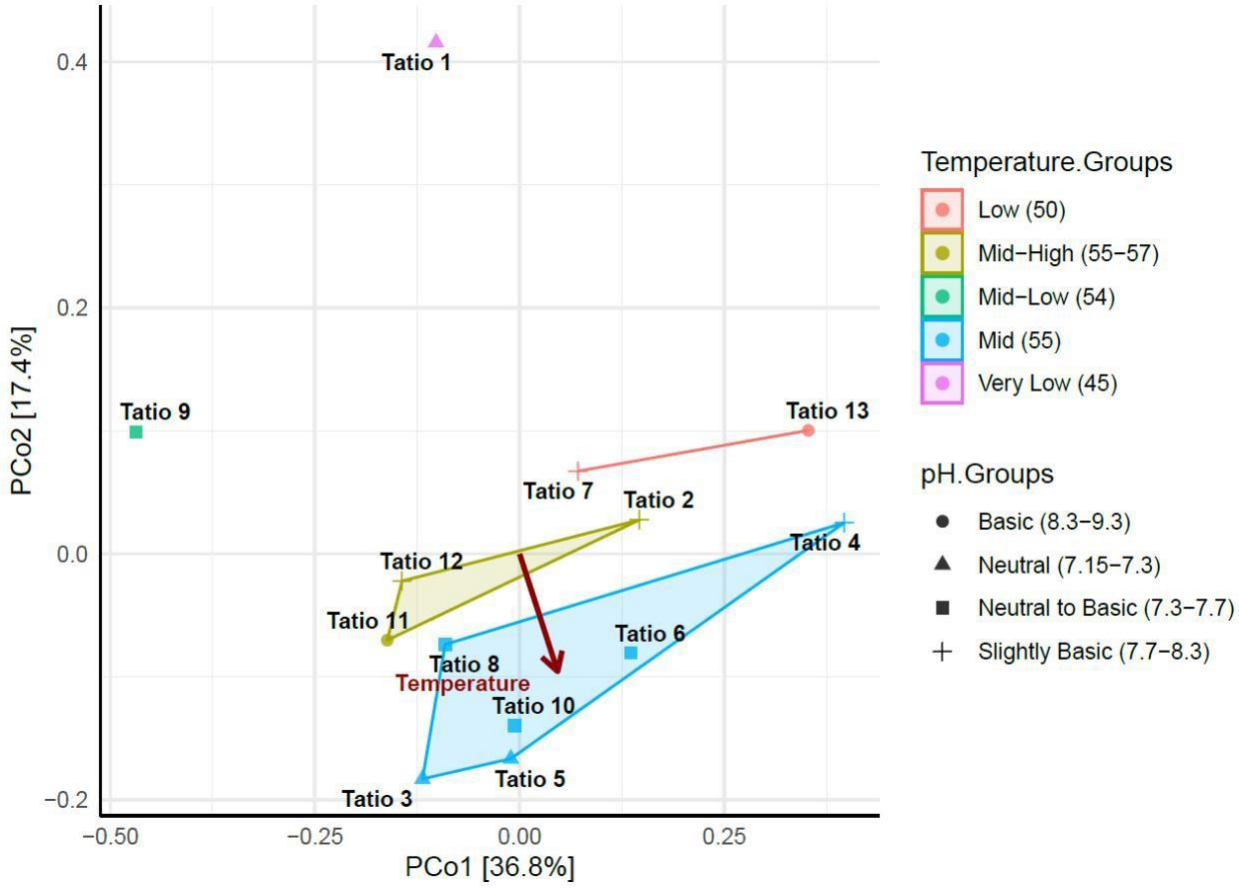

Figura 6.- Análisis de Coordenadas Principales de diversidad beta (disimilitud de Bray-Curtis) para 13 metagenomas en El Tatio.
El PCoA provee una representación de un grupo de objetos donde su relación está medida por algún índice de disimilitud (muestras similares se acercan, muestras disimilares se alejan). Se realizó una transformación de raíz cuadrada (sqrt) de la disimilitud de Bray-Curtis calculada para el grupo de metagenomas utilizados para convertirse en métrica y ser utilizada en un PCoA. Los puntos se colorearon según el rango de temperatura al que pertenecían mientras que la forma se configuró de acuerdo con su rango de pH. Los polígonos formados agrupan los grupos de temperatura. La flecha roja indica a la temperatura como un factor ambiental ajustado significativamente a la ordenación, indicando en su dirección el aumento de la temperatura.

6.2 Características de las comunidades virales activas en tapetes microbianos de termas de El Tatio y sus principales hospederos.

6.2.1 Identificación de los principales vOTUs activos y su abundancia en los tapetes microbianos de Termas en El Tatio.

Para la misma selección de sitios de muestreo dentro de los géiseres de El Tatio (13 fuentes termales distintas), se realizaron extracciones de ARN directamente desde muestras de tapete microbiano preservadas con el reactivo RNAlater, para posterior secuenciación de metatranscriptomas mediante Illumina Novaseq 6000 (Roy J. Caver Biotechnology Center, Univesity of Illinois, USA).

Luego de ser trimadas las lecturas obtenidas de la secuenciación se procedió a mapear el set de vOTUs obtenidos desde los metagenomas de El Tatio contra las lecturas forward y reverse de cada metatranscriptoma (13), para cuantificar la abundancia y actividad de los genomas virales rescatados. Se utilizó bowtie2 (Langmead y Salzberg., 2012) para el reclutamiento, y samtools para el post-procesamiento y análisis de los resultados obtenidos.

El reclutamiento del set de vOTUs representativos de El Tatio con los 13 metagenomas que lo originan, arrojó un total de 169 vOTUs correspondientes a los más abundantes (>1%) dentro de sus lecturas. Mientras que, para el caso de los metatranscriptomas, fueron 70 los vOTUs que pertenecen a esta categoría al ser mapeados contra las lecturas correspondientes (Figura 7).

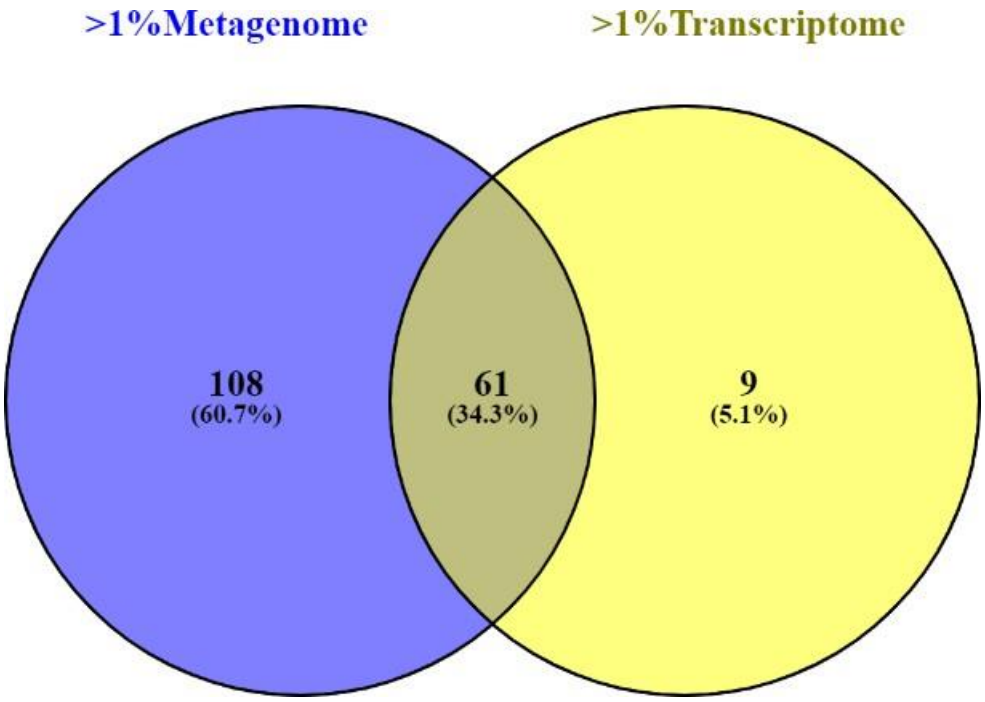

Figura 7.- Diagrama de Venn representando los vOTUs abundantes (>1%) que fueron obtenidos para las lecturas metagenómicas y metatranscriptómicas.

A partir del set correspondiente a 527 vOTUs provenientes de El Tatio, 169 vOTUs superan el umbral del 1% del total de las lecturas dentro de los 13 metagenomas, mientras que 70 cumplen este requisito en las lecturas metatranscriptómicas. De estas agrupaciones, 61 vOTUs se comparten entre ambos tipos de secuenciación, mientras que, por un lado, 108 vOTUs representan >1% de abundancia en metagenomas, no así en sus respectivos metatranscriptomas. Asimismo, 9 vOTUs tienen alta representación en la información metatranscriptómica, pero no presentan tal representatividad en sus metagenomas.

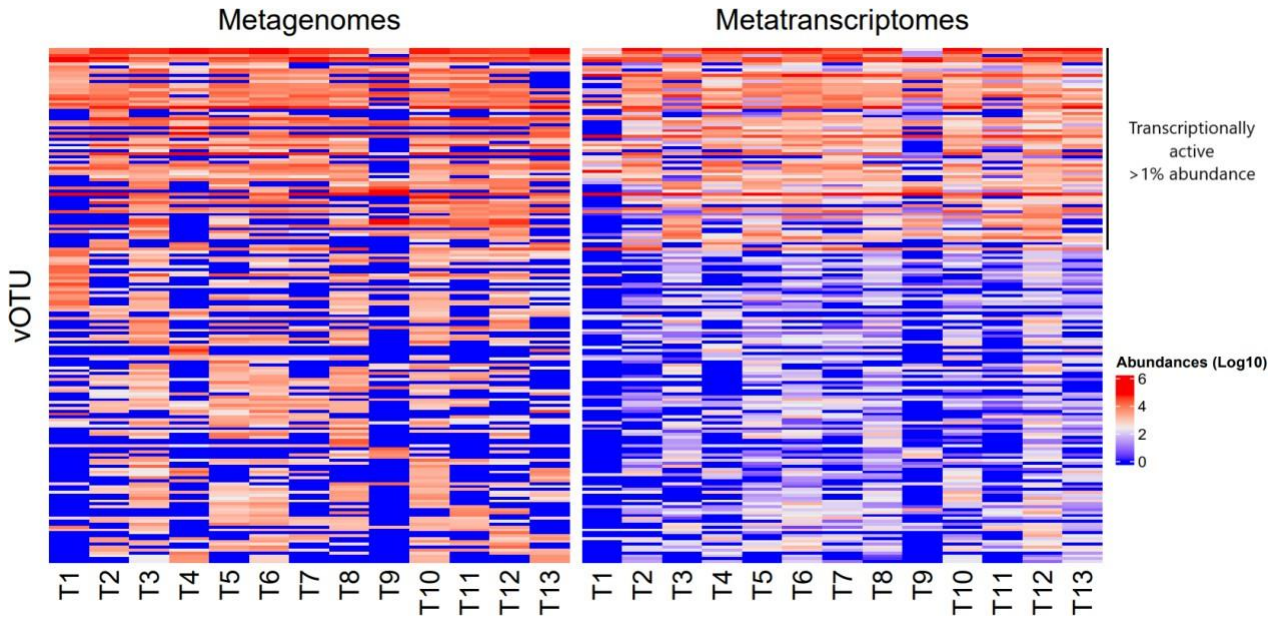

Figura 8.- Mapas de calor representando la abundancia absoluta de los vOTUs más abundantes (>1%) pertenecientes a El Tatio en los metagenomas (Izquierda) y su abundancia ajustada a las lecturas de los metatranscriptomas (derecha).

A partir de los 527 vOTUs obtenidos desde 13 metagenomas en El Tatio, se calcularon los vOTUs más abundantes dentro de los 13 sitios distintos de muestreo (T1-T13) tanto para las lecturas metagenómicas como metatranscriptómicas. En la figura 8 se representan verticalmente los vOTUs más abundantes (>1%), visualizándose su abundancia en función de un gradiente de colores (mayor presencia atribuida al color rojo y menor al color azul). A la izquierda el heatmap denota la distribución de dichos vOTUs en los metagenomas de las distintas muestras, mientras que, a la derecha, los mismos vOTUs son representados según su abundancia en los metatranscriptomas respectivos, siendo detallado en la parte superior cuáles de ellos corresponden a los más abundantes dentro de dicha data metatranscriptómica.

Una vez obtenidas las tablas de conteos normalizadas para los vOTUs transcripcionalmente activos (en este caso, no se realizó un filtró de cobertura de los vOTUs dentro de las lecturas limpias de los metatranscriptomas de cada sitio de muestreo), se procedió a observar los vOTUs más abundantes transcripcionalmente activos y compararlos con los vOTUs más abundantes en los metagenomas (Figura 8).

Estos vOTUs abundantes hacen un total de 178 secuencias que fueron representadas en los dos heatmaps concatenados (unidos horizontalmente) que verticalmente se representan para ambos casos a los mismos vOTUs (a la misma altura, un rectángulo (o fila) por cada vOTU). De esta manera, se observa que los vOTUs más abundantes (178 vOTUs en total) en gran parte también estarían siendo los más transcripcionalmente activos, evidenciándose con claridad el patrón entre la parte superior e inferior al comparar el heatmap representante de los metagenomas y metatranscriptomas (Figura 8). Es destacable que un número considerable de vOTUs se encuentran con gran abundancia en los metagenomas, pero que no estarían demostrando gran actividad transcripcional (mitad inferior de la figura 8). Sin embargo, estos vOTUs no se encuentran tan repartidos a lo largo de los 13 sitios estudiados, sino que se presentan en algunos sitios solo donde tienen temperaturas más parecidas. Algo similar estaría ocurriendo con los vOTUs abundantes en metatranscriptomas a lo largo de los 13 sitios pero que no se encuentran en los metagenomas con la misma abundancia en algunas fuentes termales específicas (Figura 8).

6.2.2 Identificación taxonómica de los principales hospederos putativos de las comunidades virales activas presentes en termas de El Tatio.

Una vez que se determinaron los principales vOTUs transcripcionalmente activos tras su reclutamiento en cada uno de los metratranscriptomas de El Tatio y se calcularon sus abundancias, se utilizaron dichas secuencias (vOTUs) y genomas ensamblados desde metagenomas (MAGs) provenientes de los mismos 13 metagenomas de estudio para tratar de determinar los posibles hospederos a los que los virus podrían estar infectando. Para esto, se recurrió a la herramienta VirMatcher 0.3.3 (Bolduc y Zayed., 2020), que predice relaciones virus-hospedero al buscar espaciadores CRISPR en el hospedero, determina profagos integrados, genes de ARN de transferencia (ARNt) y frecuencias k-mer entre los vOTUs en este caso de El Tatio y los MAGs correspondientes al set de metagenomas de trabajo. Afortunadamente ya se contaban con los MAGs provenientes de El Tatio, que fueron verificados y asignados taxonómicamente utilizando la GTDB-tk (Chaumeil et al., 2020), conjunto de herramientas que permite clasificar genomas con la base de datos de taxonomía (GTDB) para bacterias y arqueas. Con esta metodología, para 23 de los 70 vOTUs que corresponden a los más abundantes y transcripcionalmente activos, se identificaron sus hospederos putativos, al cumplir con un score de 2 o más. En general, se tiene precaución con secuencias que presenten menos de 1.5 de score o menos de dos estrategias para la asignación de hospedero.

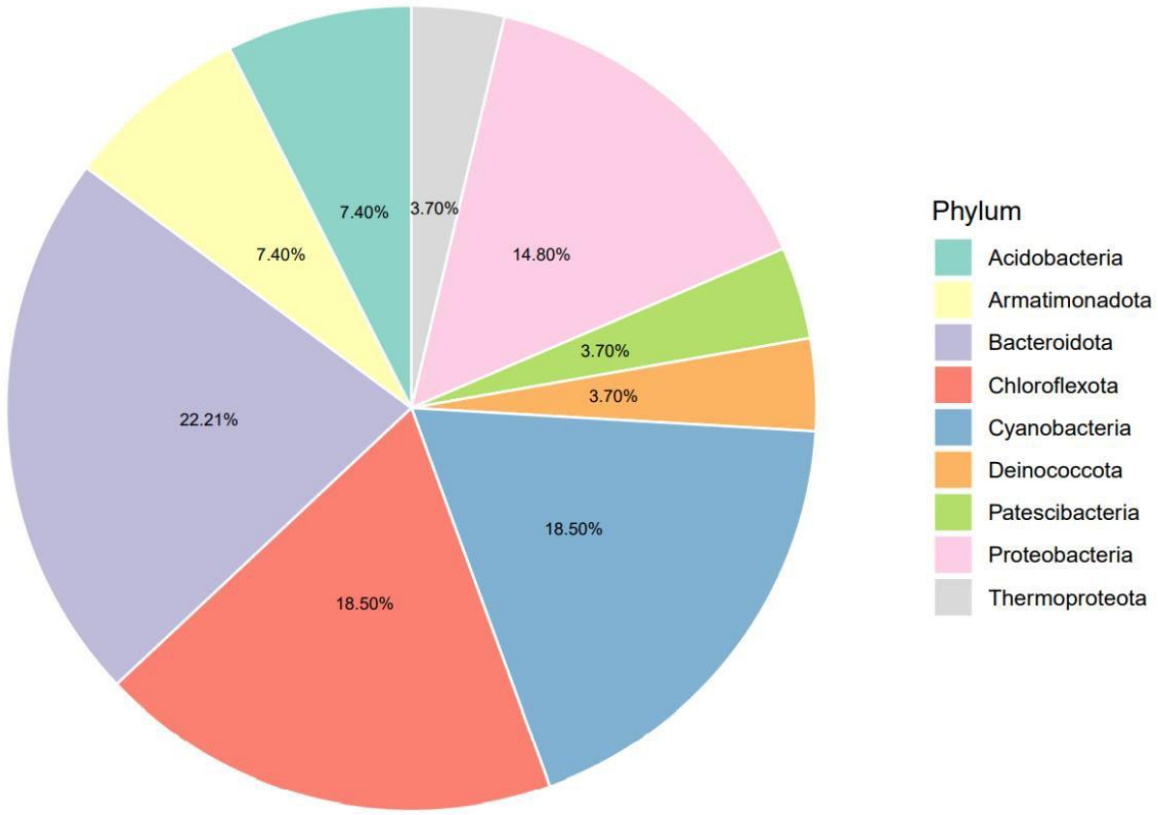

Figura 9.- Afiliación taxonómica de los principales hospederos putativos en El Tatio para los vOTUs más activos en los 13 metagenomas.
Gráfica de pastel (Piechart) representando los hospederos putativos, a nivel de phylum, de los vOTUs más abundantes a lo largo de los 13 sitios de El Tatio, tanto en sus metagenomas como sus metatranscriptomas. Para 23 de los 70 principales vOTUs transcripcionalmente activos se logró asignar a MAGs provenientes de El Tatio con la confianza suficiente para establecer la interacción virus-hospedero entre vOTU-MAG mediante VirMatcher.

En este caso, gran parte de las secuencias fueron asignadas con sólo una herramienta (WIsH), por lo que se decidió subir el umbral del score a 2.

La asignación de hospederos resultó en una predominancia de los filos *Bacteroidota* (22.21%), *Chloroflexota* (18.50%) y *Cyanobacteria* (18.50%). Seguido de filos recurrentes en estos ambientes como son *Proteobacteria* (14.80%), *Acidobacteria* (7.40%) y *Armatimonadota* (7.40%). Finalmente, en un menor porcentaje se presentan los filos *Deinococcota*, *Patescibacteria* y *Thermoproteota* (3.7%). Estos resultados que se muestran en la figura 9, demuestran que la mayoría de los virus que se estarían constantemente transcribiendo, infectan mayoritariamente a bacterias, y únicamente un grupo muy pequeño a fagos que estarían infectando Arqueas (*Armatimonadota*) (Figura 9).

Finalmente, también se obtuvo que hubo ciertos vOTUs abundantes y activos (9%) que infectan a más de un filo de bacterias con el mismo puntaje de confianza, por lo que se les consideró más de un hospedero para la asignación taxonómica de hospedero (Figura 9). La potencial promiscuidad de estos virus (9%) sugiere a nuevos filos cómo potenciales hospederos para los principales virus activos en El Tatio, resultando así también la inclusión del filo *Patescibacteria* y *Deinococcota*. Además, estos resultados evidencian también la posibilidad de que algunos vOTUs podrían estar infectando paralelamente *Cyanobacteria* y *Bacteroidota*, además de *Chloroflexota* y *Bacteroidota*.

6.3 Composición y diversidad de las comunidades de fagos de El Tatio y de otras fuentes termales de distinta fisicoquímica y ubicación geográfica en el mundo.

6.3.1 Identificación de secuencias virales desde el catálogo de contigs de los metagenomas provenientes de distintas fuentes termales del mundo. Se utilizó la misma metodología aplicada para la obtención de secuencias virales desde los metagenomas de El Tatio para el análisis de secuencias de sistemas termales a nivel global. Resumiendo, se inició el trabajo a partir de lecturas que fueron previamente ensambladas por nuestro grupo de investigación (con los mismos parámetros utilizados en El Tatio), a partir de los contigs de metagenomas seleccionados (36 metagenomas provenientes de 4 continentes y 7 países) para este estudio global. Estos metagenomas sumaron junto a los del Tatio, un total de 49 metagenomas. Este set global de metagenomas representan sistemas termales de distintas ubicaciones geográficas a lo largo del mundo, dentro de un rango de temperatura de 40- 80°C y de pH de 6-9 aproximadamente (Figura 10).

Este rango moderadamente hipertermofílico (Zablocki et al., 2017) y de pH relativamente circumneutral, permite el crecimiento de tapetes microbianos fototróficos en las fuentes termales, así como la exclusión de las principales arqueas hipertermofílicas y procariotas acidofílicos (Andersson y Banfield., 2008).

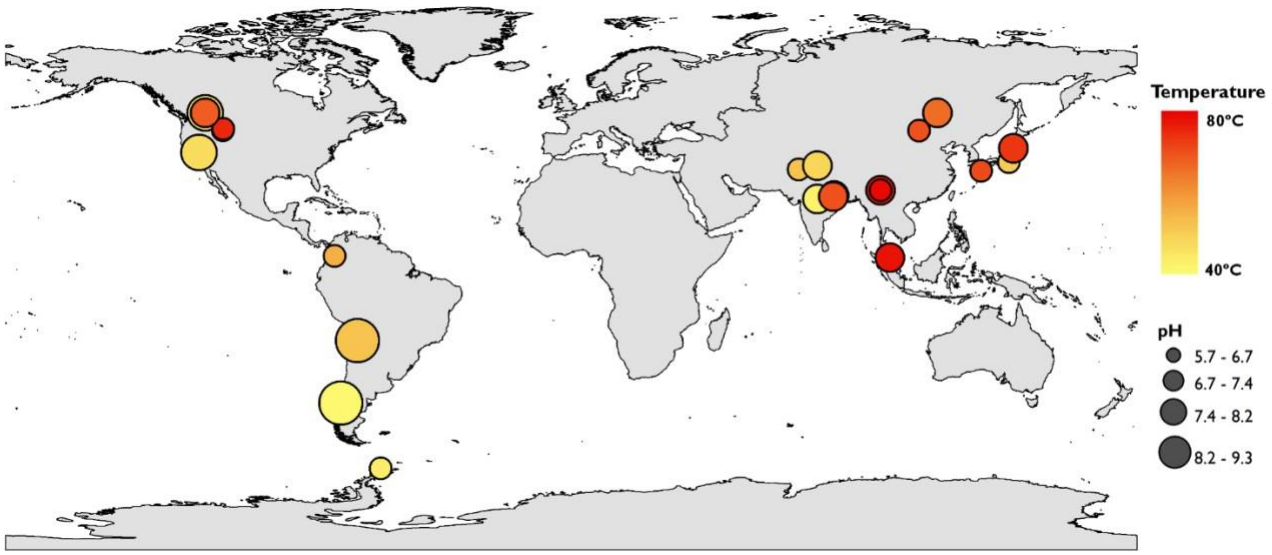

Figura 10.- Mapa global representando las ubicaciones geográficas de los 49 sistemas termales utilizados en este estudio y su principal metadata. Distribución geográfica de las termas utilizadas en este estudio, representándose el pH de cada sitio por tamaños proporcionales de círculo para cada muestra (o grupo de muestras), y gradientes de colores para determinar su posición dentro del rango de temperatura.

Para la identificación de las secuencias virales, se procedió de la misma manera que se hizo para los 13 metagenomas de El Tatio, donde WtP (Marquet et al., 2022) fue capaz de detectar en un principio un total de 10.651 contigs de procedencia viral en el total de los 49 metagenomas. Luego, tras someterse al mismo filtro de calidad descrito y realizado por CheckV (Nayfach et al., 2021), se redujo el set de contigs virales a 4.029, los que finalmente fueron clusterizados con los parámetros de CD-HIT (95% de identidad, 80% de cobertura) (Fu et al., 2012), obteniéndose el set no redundante de un total de 3.937 vOTUs.

6.3.2 Identidad taxonómica de vOTUs de sistemas termales provenientes de diversas partes del mundo.

La afiliación taxonómica de los vOTUs correspondientes al set global de metagenomas se realizó con el mismo resultado de vConTACT2 que se aplicó para la asignación de taxonomía para El Tatio. Este análisis ya contenía a todos los vOTUs provenientes de las distintas fuentes termales del mundo, con la intención de lograr la mayor cantidad de asignaciones al fomentar la formación de la mayor cantidad de VCs posibles. Sin embargo, en esta ocasión, aumenta la cantidad de asignaciones por metagenomas, ya que ahora la tabla de conteos utilizada para diseñar las gráficas de composición también va a considerar vOTUs asignados taxonómicamente que provengan de distintos metagenomas pero que posean conteos en los metagenomas de El Tatio, aportando a la composición taxonómica de estos últimos y viceversa. Al aumentar el pool de muestras a utilizar, también se modifica la normalización de dichos conteos, por lo que la abundancia relativa se verá modificada por haber un espacio muestral

mucho mayor.

Antes de normalizar las tablas de conteos por TPM, los vOTUs se filtraron por cobertura, seleccionándose los que presentaban una cobertura mayor al 75% dentro de las lecturas metagenómicas. Esto condujo a finalmente un total de 3.559 vOTUs como set global no redundante, de los cuáles 1.446 vOTUs lograron ser asignados taxonómicamente al menos a nivel de reino. Una vez tabulada esta información, se realizan las gráficas de composición a nivel de clase (Figura 11), orden y familia (Figura 12).

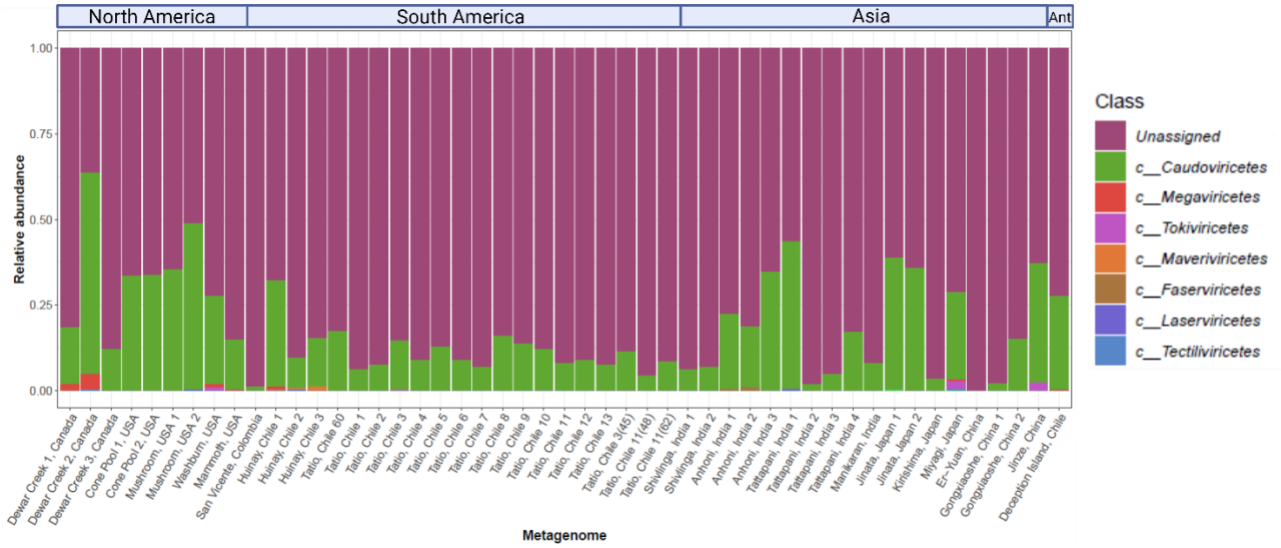

Figura 11.- Afiliación taxonómica y abundancia relativa de los vOTUs provenientes del set global presentes en los 49 metagenomas a nivel de clase.

Gráfico de composición de barras apiladas que representa la abundancia relativa de los vOTUs correspondientes a 49 metagenomas alrededor del mundo que fueron asignados taxonómicamente a nivel de clase (incluyendo las secuencias que no lograron ser asignadas). La asignación taxonómica se realizó mediante redes de genes compartidos utilizando los vOTUs de los 49 metagenomas más la base de datos IMG/VR versión 4 (2022), para ampliar en lo posible el número de asignaciones a rescatar desde vConTACT2. Las muestras están agrupadas según su procedencia geográfica.

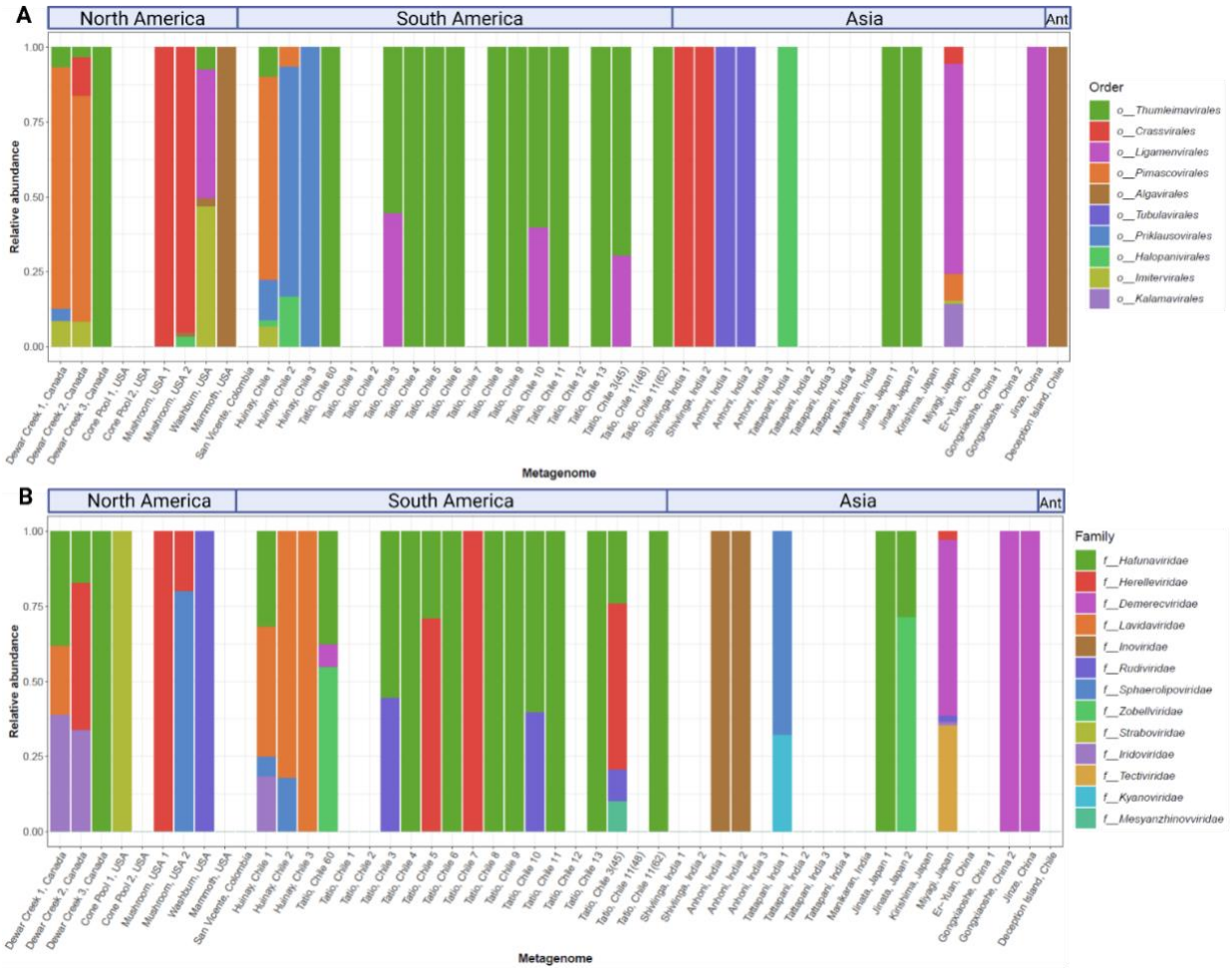

Figura 12.- Asignación taxonómica de vOTUs del set de muestras globales, presentes en cada uno de sus metagenomas a nivel de orden y familias virales. Gráficos de composición de barras apiladas representando la abundancia relativa de los vOTUs provenientes de 49 metagenomas de distintas ubicaciones en el mundo que fueron asignados taxonómicamente mediante redes de genes compartidos entre 3559 vOTUs y la base de datos del IMG/VR v4 (2022) a través del programa vConTACT2, mostrándose solamente las secuencias que lograron ser asignadas y siendo ordenados los sitios según ubicación geográfica continental. (A) Afiliaciones taxonómicas por metagenoma a nivel de orden viral, siendo representado por 62 vOTUs que lograron ser asignados con vConTACT2 a este nivel. Los metagenomas que no presentan vOTUs asignados taxonómicamente, no fueron coloreados. (B) Afiliaciones taxonómicas a nivel de familia. 47 vOTUs se lograron asignar hasta este nivel. No asignados siguen el mismo criterio.

A partir de los 1446 vOTUs asignados en el set de vOTUs no redundantes globales, todos llegaron a ser asignados hasta el nivel de clase viral, predominando casi de manera absoluta los *Caudoviricetes*, seguido de los *Megaviricetes* y *Tokiviricetes*, pero en muy bajas cantidades y mayoritariamente en termas de Norteamérica y Asia. La mayoría de los metagenomas presentan un porcentaje de secuencias asignadas menor al 50% (sólo un metagenoma de Dewar Creek, Canadá es la excepción con un ~64% asignado). A su vez, 15 metagenomas presentan una asignación mayor al 25% del total por metagenoma, siendo estas muestras pertenecientes principalmente a las regiones asiáticas y norteamericanas (sólo una en Sudamérica), evidenciando que en las termas de Chile un menor porcentaje de vOTUs fue asignado taxonómicamente respecto de otros continentes, sugiriendo una alta novedad taxonómica en esta región (Figura 11).

A nivel de Orden, fueron asignados 33 metagenomas, perteneciendo la mayor parte de los metagenomas no asignados al continente asiático (9), seguido de Sudamérica (6) y finalmente Norteamérica (2). Asia y Norteamérica presentan una distribución relativamente pareja de órdenes virales con relación al número de muestras. En Norteamérica aparece el orden *Pimascovirales* (Canadá), *Crassvirales* (Estados Unidos) e *Imitervirales* (Canadá y Estados Unidos), en Asia los órdenes *Tubulavirales* (India), *Crassvirales* (India) y *Ligamenvirales* (China y Japón). El continente antártico presenta una clara dominancia del orden *Algavirales*. En Sudamérica, representado particularmente por El Tatio (Chile), se encontró una evidente predominancia del orden *Thumleimavirales*,

destacándose del resto por la casi exclusividad de dicho orden en esta región. Sólo en algunas termas de El Tatio, aparecen también *Ligamenvirales* pero en menor cantidad, miembros que por otro lado también aparecen en minoría en el resto de los continentes (Figura 12a). Los distintos ordenes virales en Sudamérica corresponden a *Imitervirales*, *Priklausovirales*, *Halopanivirales* y *Pimascovirales*, también presentes en otras muestras diferentes a El Tatio correspondientes a la región de Huinay en la Patagonia chilena (Figura 12a). Finalmente, a nivel de familia se lograron asignar secuencias virales para 31 metagenomas del total global, 7 en Norteamérica, 15 en Sudamérica y 9 en Asia. En Norteamérica predominan con presencia similar las familias *Hafunaviridae* y *Herelleviridae*, seguidas de las familias *Iridoviridae*, *Sphaerolipoviridae*, *Straboviridae*, *Rudiviridae* y *Lavidaviridae*. En Sudamérica aparece una mayor cantidad de familias, incluyendo todas las previamente mencionadas, excepto *Straboviridae*, pero sumándose además las familias *Demerecviridae*, *Zobellviridae* y *Mesyanzhinovviridae* (esta aparece únicamente en Sudamérica). A diferencia de los patrones taxonómicos observados en Norteamérica, en Sudamérica se aprecia una predominancia evidente de la familia *Hafunaviridae,* siendo dominante en El Tatio dominado, así como gran variedad de otras familias en los metagenomas correspondientes a la Patagonia chilena Huinay (representado en 2 metagenomas) y El Tatio (un metagenoma de la Terma 3, 45°C) *(*Figura 12b).

Finalmente, para el continente asiático, se asigna de manera exclusiva las familias *Tectiviridae*, *Kyanoviridae*, *Inoviridae* y *Demerecviridae*, siendo esta

última de las más predominantes (y que en Sudamérica es muy escasa y no aparece en Norteamérica). El resto de las familias detectadas aparecen en su mayoría en Asia, aunque sea en muy baja cantidad, exceptuando *Straboviridae* (que aparece únicamente en la región norteamericana) y *Mesyanzhinovviridae* (que aparece exclusivamente en Sudamérica) (Figura 12b).

6.3.3 Diversidad de las comunidades virales en El Tatio y otras fuentes termales alrededor del mundo según sus factores ecológicos.

De manera similar a lo que se hizo con el grupo de muestras pertenecientes a El Tatio, se realizaron análisis de diversidad alfa y beta procediendo con la misma metodología, para el grupo de muestras globales.

A nivel de alfa diversidad, se aplicaron los mismos índices de diversidad (Shannon, Pielou, Chao1 y Simpson), obteniéndose en esta ocasión una distribución no normal tras la prueba de Shapiro-Wilks. Por lo tanto, se tuvieron que ajustar los datos a distintas distribuciones y elegir la que más se adecuara a ellos. Esta selección se hizo acorde a Criterios de Información de Akaike y Bayesiano (AIC y BIC, respectivamente), junto a Bondad de ajuste (Kolmogórov-Smirnov, Cramér-von Mises y Anderson-Darling), eligiéndose la distribución que presentara los valores más pequeños, para realizar los GLMs. El modelo que mejor se ajustó a los datos fue la distribución gamma.

Posteriormente, los análisis arrojaron diferencias significativas para la altitud, longitud y latitud (p-values 0.01953, 0.0286 y 0.0006549, respectivamente) según el índice de riqueza Chao1 (también conocido como Observed o riqueza observada) (Figura suplementaria 4). Gráficas de siluetas para las variables

evaluadas también determinaron que el número óptimo de clústeres para el nuevo set de datos corresponde a 2 o 3 por igual (Figura suplementaria 5). Tras realizar discretizaciones para ambas cantidades de clústeres en las variables ambientales según k-means, se obtuvieron diferencias significativas para las mismas variables observadas en los GLMs. Sin embargo, esta estrategia también reveló diferencias significativas en las comunidades virales según la variable categórica hábitat del sistema termal para la riqueza observada (Figura 13).

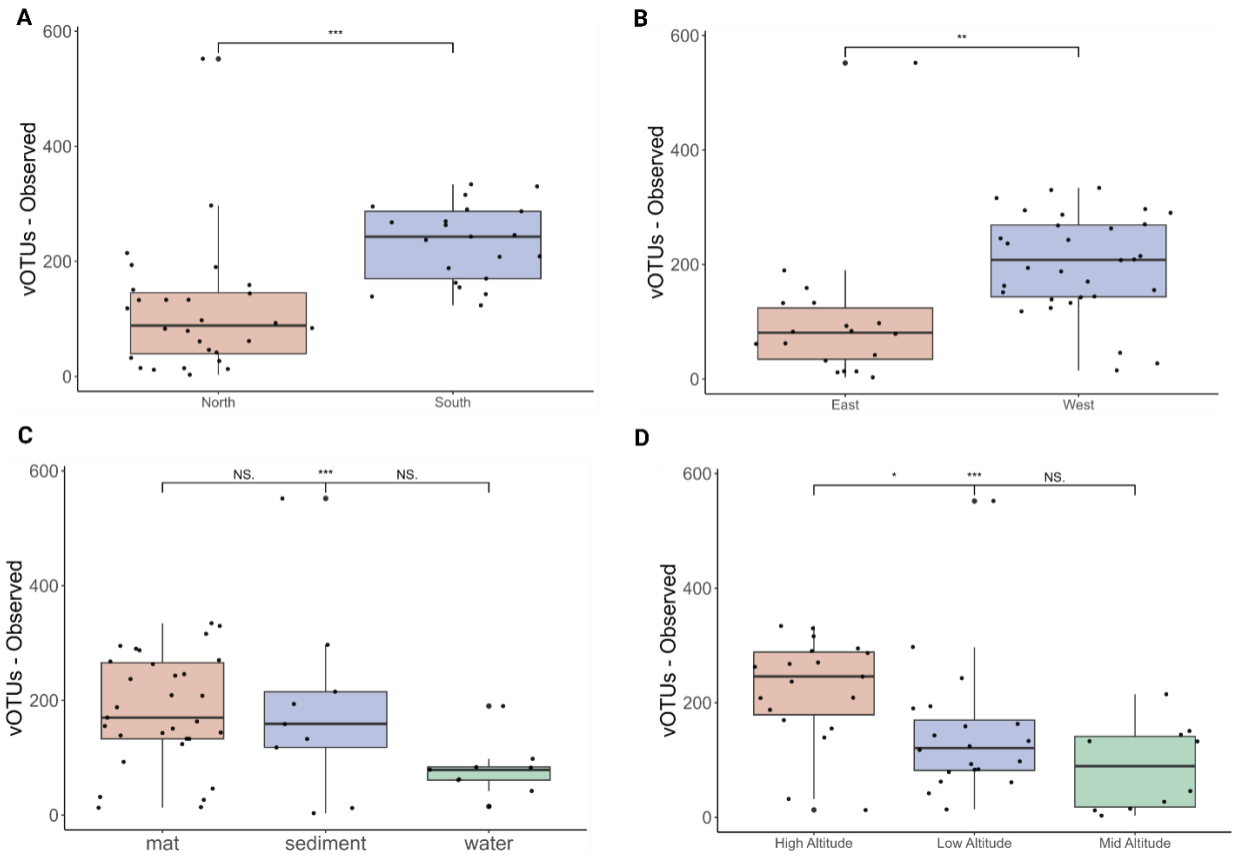

Figura 13.- Diagramas de cajas para los valores discretizados de latitud, longitud, altitud y hábitat, para los 49 metagenomas globales, utilizando el índice de diversidad Riqueza observada (Observed).

(A) Gráfico de cajas aplicado a la discretización de la latitud para las muestras globales, según su ubicación en los hemisferios norte o sur, de acuerdo con la riqueza observada como índice de diversidad alfa, obteniéndose un p-value <0.001. (B) Gráfico de cajas aplicado a la discretización de la longitud para las muestras según su ubicación entre los hemisferios oriental u occidental, de acuerdo con la riqueza observada como métrica de diversidad alfa, con un p-value < 0.001. (C) Gráfico de cajas aplicado a la variable categórica hábitat, obteniéndose diferencias significativas entre las muestras provenientes de tapetes microbiano y agua termal, con un p-value = 0.03. (D) Gráfico de cajas para la discretización de la altitud de los sitios de muestreo en 3 grupos: altitud baja, media y alta, de acuerdo con la riqueza observada es especies, con un p-value de 0.02 entre la altitud alta y baja, y < 0.001 entre la altitud alta y media.

Los resultados de alfa diversidad indican que, según la ubicación geográfica del grupo muestral a nivel continental, la riqueza de especies es mayor en el hemisferio sur y oeste de manera significativa (Figura 13a y 13b). Además, también se evidenciaron diferencias significativas para la riqueza de especies según la altitud y hábitat, revelando un aumento de esta riqueza a mayor altitud y en las muestras de tapete microbiano (Figura 13c y 13d).

Posteriormente, se procedió con un análisis de *forward selection* para determinar los parámetros que presentan diferencias significativas para la diversidad beta entre las distintas muestras utilizadas en este set global de metagenomas termales, basándose en un modelo de distancias de Bray-Curtis. Esto se aplicó para las variables continuas (Temperatura, pH, Altitud, Latitud y Longitud), mostrando como parámetro significativo a la latitud, longitud y altitud (p-value-adj = 0.010, 0.048 y 0.010, y R2-adj = 23%, 24% y 16%, respectivamente).

Sin embargo, para complementar e incluir también las variables categóricas, se realizó un PERMANOVA (999 permutaciones) basado en distancias de Bray-Curtis, donde la fuente del ADN, coordenadas (UTM), temperatura y pH, fueron incluidas al modelo de manera no secuencial e independiente.

Este análisis de varianza permutacional indicó que la temperatura, pH, coordenadas, altitud y fuente de ADN son parámetros que representan diferencias significativas entre las muestras estudiadas. De ellos, las coordenadas en UTM explicaron un porcentaje de la varianza elevado (52% de la varianza), mientras que la temperatura, pH, altitud y fuente de ADN presentaron un porcentaje de varianza de 4%, 4%, 11% y 14%, respectivamente (Tabla

suplementaria 3).

Con esta información, se optó por realizar un análisis de redundancia (RDA) para análisis de coordenadas de beta diversidad, utilizando las matrices de disimilitud de Bray-Curtis en base a la ocurrencia y abundancia de los vOTUs dentro del set global, considerando la información obtenida desde el PERMANOVA para introducir los principales parámetros en los gráficos de coordenadas (Figura 14).

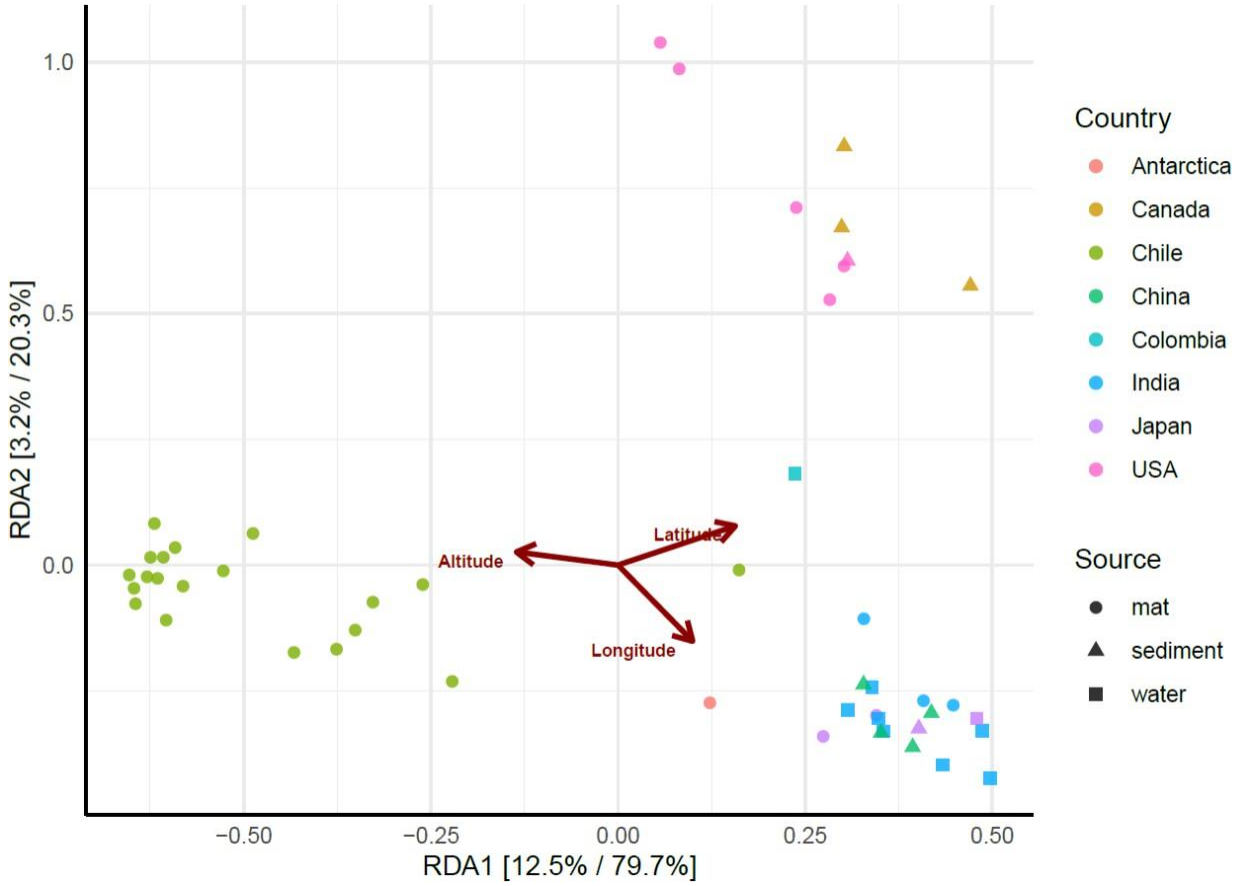

Figura 14.- Análisis de coordenadas de diversidad beta basado en matrices de disimilitud de Bray-Curtis para los 49 metagenomas de distintas localidades del mundo.

Análisis de redundancia con la latitud y longitud como variables constreñidas al ser las que presentan mayor significancia y explican el mayor porcentaje de la varianza. En este caso, a la matriz de Bray-Curtis utilizada se le hizo una transformación hellinger y se introdujeron los vectores latitud, longitud, temperatura, pH y altitud al modelo ambiental con un corte de significancia de p-value = 0.05, pasando este corte solamente la latitud, longitud y altitud, siendo visualizada su influencia con las flechas rojas y su dirección.

Se simbolizaron también las muestras según el país de procedencia (colores) y el tipo de muestra de la que se extrajo el ADN (formas).

El análisis de redundancia aplicado a los datos de ocurrencia y abundancia en conjunto a las variables constreñidas que restringen y determinan la dispersión de los datos según la varianza evidencian una clara dispersión de las muestras según su localidad, agrupándose entre ellas quienes presentan mayor similitud entre sí, resultando en 3 grupos principales que se distancian del resto, representando cada uno a un continente. Las flechas rojas denotan como la latitud determina la diferencia entre las comunidades de Norteamérica y Sudamérica, mientras que la longitud representa la distinción principalmente para el continente asiático. Finalmente, la altitud influencia principalmente a las muestras provenientes de Chile (El Tatio), debido a su extrema altitud (más de 4300 metros sobre el nivel del mar). Este resultado revela que las comunidades virales de los sistemas termales se parecen entre sí a lo largo de su misma área regional, independientemente del tipo de muestra extraída (tapete, sedimento o agua), rigiendo su disimilitud mayormente según la ubicación geográfica y no tanto o solo por propiedades fisicoquímicas como pH o temperatura, reforzando la idea de patrones biogeográficos para las comunidades virales en sistemas termales.

7 Discusión

Los sistemas termales son ambientes discontinuos que pueden ser considerados como “islas calientes” rodeadas por un “océano frío”, proveyendo un modelo de estudio único para evaluar la relevancia de los factores ambientales y limitaciones de dispersión en el establecimiento y desarrollo de comunidades microbianas (Inskeep et al., 2013; Klatt et al., 2013; Sharp et al., 2014; Power et al., 2018l).

Sumado a esto, las comunidades microbianas de estos hábitats están usualmente dominadas por pocos filos, traduciéndose en una relativa simplicidad que permite su uso para correlacionar funciones genómicas con parámetros ambientales y para entender cómo pueden determinar la estructura de la comunidad (Inskeep et al., 2010, 2013; Alcamán- Arias et al., 2015, 2018; Alcorta et al., 2018).

La composición y diversidad de las comunidades virales, particularmente de bacteriófagos, ha sido estudiada en algunos sistemas termales (Zablocki et al., 2018; Jarret et al., 2020), donde se ha demostrado que estos virus son ubicuos, numerosos y activos (Bolduc et al., 2012; Menzel et al., 2015; Munson-Mcgee et al., 2018; Schoenfeld et al., 2008; Sharma et al., 2018; Zablocki et al., 2017); además de un importante impacto ecológico en tapetes fototróficos termofílicos (Davison et al., 2016; Heidelberg et al., 2009).

Sin embargo, y a pesar de la revolución en virómica que se ha llevado a cabo durante la última década, las comunidades virales termofílicas en tapetes microbianos fototróficos siguen siendo aún poco exploradas usando estas nuevas metodologías (Guajardo-Leiva et al., 2021). Y por tanto el análisis de sistemas termales todavía tiene un gran potencial para el descubrimiento de virus, donde un número sustancial de termas adicionales necesitan ser investigadas con el fin de derivar una comprensión ecológica global de la ecología viral y sus interacciones en estos ecosistemas terrestres únicos (Zablocki et al., 2017).

Actualmente, hay un vacío en los estudios a escala intercontinental de

comunidades virales en sistemas termales, siendo descrito por primera en este trabajo de tesis, el cual, también dimensiona el potencial del campo geotermal de El Tatio en Chile como un sistema modelo ideal para comprender la composición, estructura, función e interacciones bióticas de consorcios microbianos termofílicos.

7.1 Composición taxonómica de las comunidades de virus de ADN presentes en tapetes microbianos de termas en El Tatio.

Recientemente, la taxonomía de los virus ha experimentado significativos cambios, como por ejemplo la reclasificación de los virus que infectan bacterias (bacteriófagos), siendo abolido el conocido orden *Caudovirales* de las actuales bases de datos.

Por otro lado, la nueva versión (v4) de la base de datos IMG/VR fue liberada recientemente y amplía drásticamente la cantidad de genomas virales no cultivados (UViGs, por sus siglas en inglés) disponibles de diversos ambientes (incluyendo los termales), formalizando además el uso del nuevo sistema de taxonomía establecido por el ICTV. Esto generó la necesidad de realizar nuevas asignaciones taxonómicas para los virus de ambientes termales, y más aún de aquellos tan aislados e inexplorados como son los géiseres de El Tatio. Previamente, los virus presentes en tapetes microbianos fototróficos de termas de la Patagonia en Chile se asignaron principalmente al orden *Caudovirales* y familias *Myoviridae*, *Siphoviridae* y *Podoviridae* (Guajardo-Leiva et al., 2018, 2021). Sin embargo, se desconoce todavía la mayoría de la taxonomía viral asociada a estos ambientes acorde al nuevo sistema taxonómico, siendo

completamente desconocida en El Tatio en Chile.

En esta tesis, la abundancia casi exclusiva a los *Caudoviricetes*, apareciendo puntualmente la clase *Tokiviricetes* (un único vOTU proveniente de la Terma 3; Figura 3) para los metagenomas provenientes de El Tatio, es concordante con la literatura, dado que la clase *Tokiviricetes* representa a virus que infectan arqueas, cuyas condiciones óptimas para su crecimiento es un pH cercano a 4 y temperaturas hipertermofílicas (72-98°C) (Inskeep et al., 2013; Menzel et al., 2015), condiciones que no están presentes en los tapetes microbianos fototróficos de El Tatio (temperaturas entre 45-57°C y pH entre 7,15 – 9,27). Las condiciones de El Tatio son óptimas para el crecimiento de bacterias termofílicas (Barbosa et al., 2023), por lo que se espera que los virus de estas termas representen principalmente bacteriófagos, siendo la estructura de estas comunidades virales concordantes con los hospederos que habitan dichas condiciones termales (Gudbergsdóttir et al., 2016).

La distribución espacial y afiliaciones observadas en la red de genes compartidos de proteínas virales, utilizada para establecer las asignaciones taxonómicas de los vOTUs de El Tatio (Figura 2), confirman esta hipótesis. Aquí la formación de VCs de secuencias provenientes de El Tatio con las bases de datos del IMG/VR v4 de sistemas acuáticos/termales recae casi en su totalidad en UViGs representativos de *Caudoviricetes*. Sin embargo, más allá de realizar asignaciones taxonómicas de manera directa gracias a la formación de VCs de alta fidelidad, esta red representa también una distribución espacial particular llamada “Edge-Weighted Spring Embedded Layout”. Bajo este *layout* se trata a

los nodos como imanes y las conexiones entre ellos como una fuerza de atracción o repulsión en función de la similitud de sus proteínas, de tal manera que los nodos que más se parecen entre sí se atraigan y permanezcan unidos formando clústeres, mientras que a medida que aumentan las diferencias entre sí, se distanciarán más entre sí, de tal manera, que la suma de todas estas fuerzas presentes en la red sea la menor posible (Kamada y Kawai., 1988). Bajo esta lógica, se puede apreciar la forma en que los vOTUs de El Tatio forman clústeres de manera directa con los genomas virales de la clase *Caudoviricetes* (encerrado en círculos achurados) a lo largo de la red (Figura 2). Sin embargo, en más de una ocasión se aprecia también una cercana proximidad de los clústeres representativos de *Caudoviricetes* que incluyen secuencias virales de El Tatio, con clústeres integrados por la clase *Tokiviricetes* (Figura 2). Esto se podría explicar, debido al hecho de que estos virus, que si bien pertenecen a una clase, filo y reino distinto (*Tokiviricetes*, *Taleaviricota* y *Zilligvirae*, respectivamente), presentan proteínas similares a las de un bacteriófago por la morfología que poseen similar a los de cola y cabeza (“rod-shaped”) (Snyder et al., 2007), pero no hasta el punto de establecer una conexión directa (Edge o vértice) entre ellos. Esto es concordante con el resultado obtenido de que, en ningún caso para El Tatio, se llegó a una asignación taxonómica a nivel de clase u otro nivel superior mayor al 25% del total de las secuencias presentes, indicando una alta novedad taxonómica por parte de los virus presentes en los géiseres de El Tatio. A lo largo de los 13 sistemas termales de El Tatio, la mayoría de los virus identificados corresponden al orden *Thumleimavirales* y

Ligamenvirales, con sus respectivas familias *Hafunaviridae* y *Rudiviridae* (Figura 4).

Estas familias virales se han reportado como virus que infectan arqueas con una morfología del tipo siphovirus (fagos con cabeza y una larga cola no retráctil) (Liu et al., 2021; Prangishvili y Krupovic., 2012; Baquero et al., 2020). Sin embargo, el rango de temperatura (45-57°C) y pH (7,15-9,27) de los campos geotermales de El Tatio debería restringir el crecimiento y excluir a la mayoría de las arqueas hipertermofílicas (Salgado et al., 2021), permitiendo el crecimiento de bacterias termófilas y sus bacteriófagos. Estos virus con cola que infectan arqueas, se encuentran evolutivamente relacionados con los bacteriófagos con cola de doble hebra de ADN de la clase *Caudoviricetes* y análisis genómicos y estructurales han mostrado que estos fagos con cola de arqueas y bacterias presentan organizaciones genómicas similares, con genes clusterizados en módulos funcionales y compartiendo módulos de morfogénesis de viriones homólogos, incluyendo las proteínas de la cápside (Liu et al., 2021). Junto con lo anterior, la estrategia de redes de genes compartidos basadas en proteínas virales para establecer relaciones/asignaciones taxonómicas (Jang et al., 2019), podría hacer posible que esta asignación a nivel de orden y familias de arqueas sea producto de esta indistinguible morfología entre los virus con cola que infectan arqueas y bacterias (Sourrouille et al., 2022). Esto, sobre todo, ya que a nivel de comunidades microbianas determinadas por ensambles de ARNr 16S, la presencia de arqueas para los 13 tapetes microbianos fototróficos utilizados de El Tatio fue prácticamente nula (Barbosa et al., 2023; Salgado et al., en

preparación).

De igual manera, al observar la figura 4, esta muestra un aumento en la abundancia de las secuencias virales representativas de virus de arqueas, llegando casi a igualar a los bacteriófagos en sus respectivos metagenomas (3 y 10). Sin embargo, esto es debido a que la gran mayoría de los fagos que infectan bacterias no pudieron ser asignados más allá de la clase *Caudoviricetes* y frente a lo expuesto anteriormente, podría incluso ser que estos fagos asignados como virus que infectan arqueas, puedan ser realmente bacteriófagos. Esto puede ocurrir debido a que la clasificación de los niveles taxonómicos más bajos no ha sido aun completamente formalizada, y está en un constante estado de descubrimiento y refinamiento (Evseev et al., 2023).

Actualmente la clase *Caudoviricetes* comprende cuatro órdenes y 47 familias, donde la mayoría de las familias y otros niveles inferiores taxonómicos no están asignados a órdenes, los cuales solo cubren 17 familias (Evseev et al., 2023). De cualquier forma, gracias a la nueva taxonomía viral junto al constante aumento de secuencias virales en las bases de datos, particularmente en la nueva versión del IMG/VR, es que se va vislumbrando cada vez más la enorme cantidad de especies virales que realmente pueden existir dentro de un mismo rango taxonómico y las diferencias entre ellas. Esto también queda reflejado en el presente estudio en las redes de genes compartidos utilizando estas nuevas bases de datos, como en la Figura 2. Aquí se evidenció claramente una dispersión espacial mucho mayor de la clase *Caudoviricetes*, generándose muchos más clústers separados entre sí, pero que pertenecen a la misma clase

(clústeres de nodos de color rojo, Figura 2). Esto indica la alta probabilidad de que sean más bien distintos órdenes, familias y niveles taxonómicos más bajos que puede englobar una clase y que faltan por descubrir. Comúnmente estas redes usando las bases de datos previas, simplemente agrupaban en un gran clúster la mayoría de los virus que correspondían al abolido orden de los *Caudovirales* (Figura suplementaria 2).

Finalmente, notar que hay un gran número de vOTUs provenientes de las termas de El Tatio que se alejan de todos los clústeres representativos de secuencias virales referenciadas en la red (no visibles en la Figura 2), y que se agrupan únicamente entre sí formando numerosos clústeres exclusivos de El Tatio de diversos tamaños (Figura suplementaria 1). Este resultado estaría indicando una alta novedad y diversidad taxonómica que existe en este lugar, consecuencia posiblemente de las condiciones ambientales particulares y geográficas en las que se encuentra este campo de géiseres en Chile.

7.2 Diversidad de las comunidades de virus de ADN presentes en tapetes microbianos de termas en El Tatio.

Se ha sugerido que los virus constituyen el principal factor que puede moldear la diversidad de sus hospederos a través de la coevolución (Sano et al., 2018), y regular la estructura de las comunidades celulares mediante depredación en ambientes termales (Breitbart et al., 2004; Klatt et al., 2013; Pride y Schoenfeld., 2008; Schoenfeld et al., 2008). Ahora bien, aunque se han realizado investigaciones sobre el impacto ecológico y la abundancia de los virus en tapetes fototróficos termofílicos (Davison et al., 2016; Heidelberg et al., 2009),

actualmente, la diversidad de estos virus permanece siendo muy poco estudiada. Estudios en metagenómica celular en la terma Manikaran de India han analizado la diversidad viral en tapetes microbianos fototróficos, encontrándose una baja diversidad de virus (debido posiblemente a la baja cobertura de estos metagenomas), que incluye 14 genomas virales de las familias *Myoviridae* y *Siphoviridae* (Sharma et al., 2018). Por otra parte, trabajos en el campo geotermal Yellowstone (Estados Unidos), han evidenciado muy pocos virotipos compartidos entre distintos muestreos del mismo sistema termal estudiados mediante análisis de secuencias CRISPRs. Esto sugiere que la diversidad de las poblaciones virales podría o bien ser tan alta que abrumarían a los sistemas CRISPR, o que la respuesta de este sistema de defensa bacteriano a los ataques virales es muy rápida y localizada, indicando finalmente la posibilidad de que los tipos virales superan con creces en número la diversidad de especies microbianas en estos tapetes microbianos, sugiriendo una dinámica entre los fagos y bacterias de cambios muy rápidos (Heildelberg et al., 2009; Davison et al., 2016). Estas suposiciones son respaldadas también por experimentos manipulados de inducción viral de Guajardo-Leiva et al., (2021). Estos autores mediante inducción estudiaron las tendencias líticas/lisogénicas de los virus en tapetes microbianos fototróficos, proponiendo que para estos ambientes con predominancia a la lisis en virus que infectan a los productores primarios aumentaría la diversificación mediante una evolución antagónica entre el virus y su hospedero, contribuyendo a la microdiversidad encontrada en termas.

Por otra parte, también se ha mostrado que la maduración de estos tapetes o

biopelículas y su complejidad estructural es crítica para proteger a las bacterias de un flujo continuo de fagos desde afuera del tapete (Vidakovic et al., 2018). Estudios de diversidad viral acorde a la estratificación de los tapetes microbianos han detectado una mayor diversidad viral en las capas más externas próximas a la superficie del tapete, donde se han reportado mayores conteos virales respecto del interior (Jarret et al., 2020).

En esta tesis, los trece metagenomas obtenidos desde El Tatio, se utilizaron para cuantificar la diversidad alfa viral mediante los índices de Shannon, Pielou, Simpson y Chao1. Sin embargo, tras una normalización y regresión de los datos, no se pudieron apreciar diferencias significativas de alfa diversidad a lo largo del gradiente de temperatura y pH. Sin embargo, al modelar la latitud, longitud y altitud con los mismos índices de diversidad, la latitud evidencia diferencias significativas en la alfa diversidad según la equidad de Pielou (p- value = 0.05099) y equidad de Shannon (p-value = 0.03501) entre las muestras de El Tatio. Estudios en otros tapetes microbianos de Japón han detectado que la abundancia y patrones de diversidad a la escala de centímetros a metros son explicados por el gradiente de temperatura (Miller et al., 2009; Everroad et al., 2012; Power et al., 2018) mientras que el pH y la geoquímica son más importantes a las escalas kilométricas (Fierer y Jackson., 2006; Power et al., 2018). Por consiguiente, se esperaría que para este caso aparecieran como factores significativos la temperatura y pH por sobre la latitud. Esto podría estar ocurriendo debido a la selección de muestras utilizada en nuestro estudio, donde gran parte pertenecen a una misma temperatura (6 de las 13 muestras utilizadas

presentan 55°C) y pH (7.2-7.8), y por tanto la falta de muestras que puedan aumentar la representatividad hacia las temperaturas más bajas y altas, así como muestras de pH más ácidos.

Sin embargo, el hecho de que ambas medidas de equidad presenten diferencias significativas en la alfa diversidad respecto de la latitud, indicaría que a medida que aumenta la latitud (Figura 5) estaría aumentando la equidad de especies, es decir, éstas tendrían abundancias más similares entre ellas, presentando menos especies “raras” (o poco abundantes). Al observar la Figura 1B y 1C, se pueden apreciar 3 agrupaciones de muestras al recorrer la zona geográfica latitudinalmente, este gradiente junto a los resultados de alfa diversidad estarían indicando que las muestras presentes en la parte superior de la figura 1B y 1C (Terma 1, 2, 5, 6, 7, 11 y 12) deberían ser quiénes presenten mayor equidad en sus especies. Sin embargo, resulta llamativo que este grupo, sea a su vez el más variado en parámetros fisicoquímicos: posee todo el rango de temperatura disponible (45-57°C) y un amplio rango de pH (7,3-8,6), indicando que, en esta particular zona del El Tatio, la diversidad de las comunidades virales se rige más por una alta dispersión local que por las propiedades fisicoquímicas. Estos resultados abren nuevas preguntas que evalúen posibles efectos de dispersión por otros factores como el aire ayudado por el efecto de las “erupciones” y emanaciones de vapor de agua constantes que se dan en estos géiseres. Por otro lado, en este estudio falto analizar otros parámetros fisicoquímicos diferentes como concentración de sulfuros o azufre elemental, y evaluar cómo estos podrían afectan a estas comunidades virales, ya que han sido descrito también como

factores importantes en la estructuración de las comunidades bacterianas en estos ambientes (Inskeep et al., 2010, 2013; Menzel et al., 2015).

Respecto a la diversidad beta, análisis multivariados permutados de la varianza corresponden a la longitud y temperatura como las variables ambientales que explican el mayor porcentaje de la variabilidad para la composición de las comunidades virales de los géiseres de El Tatio de manera significativa (p-value = 0,027, R2 = 19,8 %; p-value = 0,048, R2 = 17,8%, respectivamente) (Tabla suplementaria 3), evidenciándose una distribución y agrupación de las comunidades virales de acuerdo con las temperaturas de sus respectivas termas de origen, independientemente del pH que presentan (Figura 6). Junto a esto, la distribución observada en el análisis de coordenadas principales también responde a patrones geográficos, distribuyéndose con mayor cercanía las comunidades virales más próximas entre sí geográficamente (Figura 1, Figura 6). Esto indica que los virus termófilos de los géiseres de El Tatio, además de estar posiblemente regidos por la temperatura como un motor de su diversidad, la distancia geográfica estaría determinando aún más su diversidad, siendo las comunidades virales de termas con mayor cercanía más similares entre sí, independientemente de sus características fisicoquímicas. El pH no sería un factor determinante de las diferencias entre las comunidades de fagos para las termas en El Tatio (Figura 6).

En sistemas geotermales en Nueva Zelanda, se ha descrito distribuciones de diversidad beta para la comunidad microbiana hospedera de estos virus en el agua, donde se han encontrado correlaciones significativas más fuertes con el

pH que con la temperatura, siendo dicho pH el determinante de la variación observada en la diversidad entre distintos sitios del campo geotermal (Power et al., 2018). Además, este mismo estudio estableció una tendencia positiva de decaimiento de la similitud al aumentar la distancia geográfica, sugiriendo que existe una limitación de la dispersión de la comunidad microbiana entre campos geotermales distantes dentro de Nueva Zelanda, siendo despreciable la diferencia dentro de una misma terma. Sin embargo, también encontraron que la mayor diferencia entre disimilitudes de Bray-Curtis ocurria entre termas adyacentes (< 1.4 metros de distancia), siendo en estos casos la temperatura el factor que mejor correlaciona con la diversidad beta (Power et al., 2018). Por otro lado, a nivel de tapetes microbianos fototróficos, se ha caracterizado la comunidad bacteriana en termas de Costa Rica con temperaturas (37-63°C) y pH (6-7.5) similares a los de este estudio, donde análisis multivariados indicaron que el pH (por sobre la temperatura) era el primer factor en influenciar las diferencias en la composición de la comunidad bacteria, sugiriendo que la visión tradicional de que la temperatura es el principal conductor de la diversidad en los sistemas termales necesita ser revisada (Uribe-Lorío et al., 2019).

Finalmente, a nivel de comunidades virales, se ha identificado en tapetes microbianos fototróficos en termas de la Patagonia Chilena que la composición de especies virales no estaba asociada a variables ambientales como la temperatura y el pH de manera significativa, mostrando que los distintos sitios de muestreo finalmente eran los que contribuían a la varianza en la estructura de la comunidad viral (Guajardo-Leiva et al., 2021).

De esta manera, en los tapetes microbianos de los géiseres de El Tatio, las comunidades virales podrían no estar rigiéndose por sus hospederos para determinar su diversidad en la composición de especies, sino que en la ubicación geográfica de cada termal individual. Esto, debido a las significancias y varianzas contrastantes para la temperatura y el pH (en las comunidades virales) respecto de lo reportado para las comunidades bacterianas en este tipo de ecosistema (siendo particularmente contrario a lo encontrado para dicha contraparte bacteriana en el trabajo de Barbosa et al., 2023 en los géiseres de El Tatio), dándole un mayor peso a la ubicación geográfica. Por consiguiente, las comunidades virales de El Tatio reportan la posibilidad de patrones de biogeografía a escalas locales, que en parte podrían ser influenciados por la temperatura del sistema, a pesar de que se presenten los mismos filos de hospederos a lo largo de las 13 termas investigadas (Barbosa et al., 2023; Salgado et al., en preparación).

Sin embargo, aún queda por determinar si este fenómeno podría ser causado por una selección de nicho que impulsa a la comunidad microbiana en una escala local (Power et al., 2018) o debido a factores geoquímicos no evaluados en este trabajo, como sugiere Barbosa et al., 2023. De todas maneras, este trabajo ilustra la marcada heterogeneidad espacial y procesos selectivos que pueden existir entre campos geotermales individuales, reforzando la idea de los géiseres de El Tatio como un sistema modelo para poner a prueba los factores ecológicos que estructuran sus comunidades tanto microbianas como virales.

7.3 Alta actividad transcripcional de las comunidades virales más abundantes (>1%) en los tapetes microbianos de termas en El Tatio.

La actividad de los virus de tapetes microbianos fotótrofos permanece en gran parte inexplorado, existiendo muy pocos estudios que hayan proporcionado información sobre la presencia de virus activos en estas comunidades (Heidelberg et al., 2009; Davison et al., 2016; Guajardo-Leiva et al., 2018). En particular, en Chile, trabajos previos de nuestro grupo de investigación han reportado en tapetes microbianos de la terma Porcelana (Patagonia Chilena), que los virus activos dominantes pertenecen a las familias *Myoviridae*, *Podoviridae* y *Siphoviridae*, dentro del orden *Caudovirales*. Nuestros estudios en Porcelana también evidencian una leve disminución en los transcritos asociados a los *Caudovirales* a medida que aumenta la temperatura en el gradiente de esta terma, debido a una reducción de secuencias relacionadas a los morfotipos (previamente llamadas familias) Miovirus y Podovirus. Esto fue determinado al reclutar los genomas virales obtenidos con las lecturas de cada metatranscriptoma obtenido de esta terma, correspondiendo al 68% de las lecturas (Guajardo-Leiva et al., 2018). Utilizando este mismo principio, en el presente estudio de los tapetes microbianos de los géiseres de El Tatio en Chile, se encontró que las comunidades virales más abundantes, representadas por 169 vOTUs (obtenidos desde 13 metagenomas), también corresponden a los vOTUs más activos (70 vOTUs que corresponden a los más abundantes (>1%) en los metatranscriptomas) (Figura 7, 8). Una gran parte de estos 70 virus activos no son conocidos (es decir, sin asignación taxonómica) y más del 50% no se

encuentran afiliados a otras secuencias virales (singletons). No obstante, un 10% de estos virus activos que sí pudieron asignarse, pertenecen a la clase *Caudoviricetes.* Lo esperado es que el resto de vOTUs también corresponda a esta clase *Caudoviricetes* o que presenten una gran cercanía a miembros de esta clase (previamente catalogada como el orden *Caudovirales*) ya que típicamente infectan bacterias y algunas arqueas no hipertermofílicas (Maniloff y Ackermann, 1998). Por su parte, en estos microambientes de los tapetes microbianos termales, los hospederos que comúnmente predominan son las cianobacterias fototróficas oxigénicas y los fotótrofos anoxigénicos filamentosos del filo *Chloroflexota*, los que se ubican en las capas superiores (Alcamán-Arias et al., 2018), mientras que a niveles más profundos se encuentra una plétora de bacterias y arqueas quimioheterotróficas (Guajardo-Leiva et al., 2021). Los virus de estas biopelículas han sido estudiados principalmente a través de las asociaciones con sus hospederos (Howard-Varona et al., 2017; Guajardo-Leiva et al., 2018; Guajardo-Leiva et al., 2021), para lo que se han aislado y caracterizado fagos termofílicos con sus correspondientes hospederos (Zablocki et al., 2018), sin embargo, el sobreuso del género *Thermus* como un hospedero permisivo para el descubrimiento de nuevos fagos (basado en cultivos) ha generado un sesgo. Esto podría remediarse utilizando un mayor rango de hospederos como Cianobacterias, Chlroflexi y Aquificae (Zablocki et al., 2018).

En esta tesis, para los 70 vOTUs más abundantes y activos transcripcionalmente de los 13 metagenomas de géiseres de El Tatio (Figura 7), encontramos que solamente 2 corresponden a secuencias de prófagos. Esto es un indicador de

que en estos ecosistemas termales de El Tatio predominaría el ciclo de vida viral lítico, donde las infecciones productivas estarían directamente influenciando la composición de la comunidad microbiana a través de dinámicas depredador-presa, en la cual los taxones dominantes o activos de la comunidad microbiana estarían siendo selectivamente lisados, de acuerdo con el modelo ecológico KTW (Knowles et al., 2016). Aunque comúnmente se ha sugerido que el ciclo lisogénico es el estilo de vida viral dominante en las comunidades microbianas de termas (Sharma et al., 2018), sólo se ha demostrado experimentalmente (indirectamente) mediante secuenciación de células planctónicas provenientes de tapetes microbianos no fototróficos en condiciones hipertermofílicas (74 a 82°C) en sistemas termales de California, USA (Breitbart et al., 2004; Jarett et al., 2020), así como también recientemente en tapetes microbianos fototróficos de la terma Porcelana en Chile (Guajardo-Leiva et al., 2021). En este último, se muestra que, en los tapetes microbianos de esta terma no ácida, el ciclo lítico predomina en los grupos virales que infectan productores primarios (*Cyanobacteria*, que también corresponden al filo más activo y abundante), mientras que el ciclo lisogénico predomina en los virus que están infectando bacterias quimioheterotróficas (*Proteobacteria*, *Firmicutes* y *Actinobacteria*) (Guajardo-Leiva et al., 2021).

7.4 Hospederos putativos de los principales virus activos en los tapetes microbianos fototróficos de El Tatio.

Para corroborar la naturaleza lítica o lisogénica de cada virus detectado en El Tatio, además de la clasificación de profagos que realiza CheckV (Nayfach et al., 2021), se recurrió en este estudio al uso de la nueva herramienta geNomad, que cuenta con más de 200 mil marcadores específicos para virus y hospederos, que se analizan según proximidad para establecer provirus putativos (Camargo et al., 2023).

En el caso de los géiseres de El Tatio, los principales hospederos para los virus más activos corresponden a los filos *Cyanobacteria* (36%) y *Chloroflexota* (22%) (Figura 9), comprendiendo más de la mitad de la comunidad objetivo de infecciones virales. Estos filos, principalmente las cianobacterias, han sido reportadas previamente en los tapetes microbianos fototróficos como el taxón más importante y activo (entre 40-50 °C), siendo responsables de la mayoría de la producción primaria en este tipo de termas no ácidas y temperatura en el rango termófilo, a través de la fotosíntesis oxigénica y fijación de nitrógeno atmosférico (Alcamán-Arias et al., 2018). En estos tapetes, *Chloroflexota* puede utilizar también la luz y realizar fotosíntesis anoxigénica, además de fijar dióxido de carbono y captar amoníaco, volviéndose cada vez más activo y predominante a medida que aumenta la temperatura (Alcamán-Arias et al., 2018). De los 496 MAGs rescatados desde metagenomas de El Tatio que fueron utilizados para la predicción virus-hospedero, se encuentran representados en un 57% por el filo *Chloroflexota*, seguido de los filos *Armatimonadota* (18%), *Bacteroidota* (8%) y

Cyanobacteria (6%) (Alcorta et al., en preparación).

Transcripcionalmente, *Chloroflexota* también corresponde al filo más abundante en estos ensambles (67%), seguido por *Armatimonadota* (12%), *Acidobacteriota* (7%) y *Cyanobacteria* (5%) (Alcorta et al., en preparación), indicando que gran parte de los virus más abundantes y activos también infectan a los filos bacterianos más abundantes y activos en estos tapetes. Sin embargo, el principal hospedero identificado para muchas secuencias virales corresponde al filo *Cyanobacteria*, a pesar de su baja abundancia a nivel de MAGs en metagenomas (6%) y metatranscriptomas (5%) (Alcorta et al., en preparación). La carrera armamentista evolutiva entre pares cianofágo-cianobacteria específicos en ambientes naturales es conocido por presentar escenarios donde una población viral específica se puede volver extremadamente virulenta y causar la lisis de la población del hospedero según el modelo ecológico “maten al ganador” (KTW por sus siglas en inglés) (Andersson y Banfield., 2008; Guajardo-Leiva et al., 2018). Estudios previos de nuestro grupo de trabajo en tapetes microbianos fototróficos de la Terma Porcelana (Chile) demostraron que existe una “lucha” activa constante entre un cianofago (TC-CHP58) y la cianobacteria *Mastigocladus (Fischerella)* (Guajardo-Leiva et al., 2018), donde estos cianopodovirus constituyen uno de los grupos virales más abundantes y activos transcripcionalmente de la comunidad (90% de los transcritos correspondientes al abolido orden *Caudovirales*) (Guajardo-Leiva et al., 2018; Guajardo-Leiva et al., 2021). Esta carrera evolutiva armamentista entre el par específico cianofago-cianobacteria, puede generar una mayor resistencia del hospedero que fuerce

una disminución de la población viral, o el aumento de virulencia en un grupo específico de virus, generando el decaimiento en la población del hospedero (modelo KTW) (Guajardo-Leiva et al., 2018). Si bien de los 13 metagenomas de géiseres de El Tatio hay 10 que fueron obtenidos a temperaturas mayores o igual a 55°C (Tabla suplementaria 1) (temperatura desde la cual se ha reportado también para tapetes microbianos del Parque Nacional Yellowstone (USA) que la cianobacteria *Fischerella thermalis* ya no crece en su óptimo (Brock y Brock., 1966)), se ha descrito que en tapetes microbianos fototróficos de Porcelana de similar rango de temperatura y pH a los de El Tatio, el estilo de vida viral lítico sería el dominante en los productores primarios, como es la cianobacteria dominante *Fischerella*, encontrándose que uno de los grupos virales líticos más abundantes corresponde a los cianofágos, que infectarían a esta cianobacteria *Fischerella* (Guajardo-Leiva et al., 2021). Otros estudios han confirmado que la temperatura límite de crecimiento para esta bacteria en cultivo se encuentra entre 57-58°C (Schwabe 1960; Muster et al., 1983; Milleret al, 2007; Finsinger et al., 2008; Alcamán et al., 2017; Vergara-Barros et al., 2022), aunque el límite superior de temperatura puede variar dependiendo de la geoquímica original de la terma (Schwabe, 1960).

Experimentos previos de lisis en cianobacterias provenientes de aguas dulces demostraron que, en 4 días, una comunidad activa disminuye sus niveles de clorofila α a un 10% de sus niveles máximos, además de evidenciar mediante microscopía electrónica fagos del tipo lambda tanto en el medio de cultivo como adheridos a los filamentos de las cianobacterias y densamente empaquetados

dentro de los filamentos lisados de las mismas (Van Hannen et al., 1998). Paralelo a esto, a nivel de ARN ribosomal 16S, se constató que, en una electroforesis en gel con gradiente de desnaturalización, la mayoría de las bandas identificadas relacionadas a cianobacterias desaparecieron posterior al evento de lisis (Van Hannen et al., 1998).

Nuestra hipótesis, basada en observaciones y resultados propios, y su comparación con la literatura, sugieren que posiblemente existe una distribución de nicho de la comunidad bacteriana producto de la temperatura en este tipo de sistemas termales como los de El Tatio, donde se detendría el crecimiento de las cianobacterias como *Fischerella* por sobre los 55ºC, pero gran parte de su biomasa aún permanecería por un tiempo a esas temperaturas, albergando virus adheridos a sus filamentos o empaquetados dentro de los filamentos lisados, como describe Van Hannen et al (1998) y por ende, obtenerse aún un alto conteo de secuencias virales predichas que infecten al filo *Cyanobacteria*.

Finalmente, para las secuencias virales con predicciones de hospederos corresponden a los principales filos (*Chloroflexota* y *Cyanobacteria*) se obtuvo 40 vOTUs distintos que infectan a *Chloroflexota*, y 15 a *Cyanobacteria*, indicando la posibilidad de que efectivamente la alta predicción de hospederos indicada en la Figura 9 de esta tesis hacia las cianobacterias, sea producto de una reducida comunidad activa de cianobacterias, pero con alta biomasa, representando el resultado del ciclo productivo de los virus líticos replicados en pocos hospederos. Mientras, en el caso de *Chloroflexota*, los virus se encontrarían activamente produciendo una progenie viral en un rango más amplio de hospederos, al

encontrarse en un rango de temperatura óptima para su crecimiento (Feiner et al., 2015; Alcamán et al., 2018). Si bien se han descrito distintos rangos de temperatura óptima para el crecimiento de *Chloroflexota*, que oscilan entre 35-70°C (Nübel et al., 2002), resultados de ensamble de marcadores ARNr 16S mediante la herramienta MATAM (Pericard et al., 2017) para los géiseres de El Tatio corroboraron la predominancia del filo *Chloroflexota, donde* su abundancia y actividad prevalece en El Tatio con un 42% y 83% (Salgado et al., en preparación) respectivamente. Esto coincide con la representatividad observada en los MAGs aquí estudiados, destacando nuevamente la alta actividad transcripcional de este filo a lo largo de los tapetes microbianos de El Tatio. Los virus estarían, por tanto, desempeñando un papel clave en la regulación de la comunidad microbiana de estos tapetes en El Tatio, al estar activamente infectando a quiénes serían los principales productores primarios, tal como se ha reportado en otros ambientes (Kawai et al., 2021). Por tanto, la predicción virus-hospedero es sumamente importante y novedosa, ya que previamente solamente se han realizado asociaciones entre Caudovirus y *Chloroflexus* sp. usando espaciadores CRISPR-Cas (Paez-Espino et al., 2016), además de encontrarse que, en ambientes termales, el filo *Chloroflexota* es quién más se encuentra afiliado al gen *cas1* (Salgado et al., 2021), sugiriendo la interacción de este par virus-hospedero que hasta el día de hoy no ha sido demostrado.

En cuanto a los filos bacterianos que representan a la mayoría del resto de los posibles microorganismos hospederos de los virus más activos en las termas de El Tatio, encontramos a los filos *Bacteroidetes* (18%), *Armatimonadetes* (13%) y

Proteobacteria (6%) (Figura 9). En su mayoría, estas bacterias tienen un metabolismo quimioheterotrófico, encontrándose en las capas más profundas del tapete, e interactúan con los productores primarios mediante el ciclaje de nutrientes y elementos (Klatt et al., 2013; Guajardo-Leiva et al., 2021; Abed et al., 2018; KC-Y Lee et al., 2014). No obstante, su abundancia y contribución a la producción primaria, en tapetes microbianos, es relativamente baja (Alcamán-Arias et al., 2018). Se ha encontrado que algunos de estos filos son principales hospederos en tapetes y sedimentos termales en las termas de Manikaran, Estados Unidos (Sharma et al., 2018), siendo algunos casi exclusivamente asociados con fagos lisogénicos como en la terma Porcelana en Chile (Guajardo-Leiva et al., 2021).

Sin embargo, las secuencias virales correspondientes a provirus en El Tatio, representan una minoría (2,9%) del total de vOTUs obtenidos. Esto implicaría, que el modelo “a cuestas del ganador” (PTW, por sus siglas en inglés; Silveira y Rohwer., 2016), no sería representativo del comportamiento de los virus en estos ambientes en El Tatio. Este modelo describe que la abundancia del hospedero es la fuerza que rige el cambio entre lisis/lisogenia de los virus, donde un aumento en la densidad de los hospederos conlleva un aumento de la lisogenia y así conferir una ventaja competitiva a los filos dominantes, de manera que disminuye la oportunidad de crecimiento de los filos menos dominantes y así reducir la diversidad de los hospederos (Chen et al., 2021).

Ahora bien, en estos tapetes fototróficos, los nutrientes no son un factor limitante, siendo la abundancia y tasa de crecimiento de bacterias más alta que en tapetes

oligotróficos, lo que podría sugerir que la lisogenia es la dinámica favorecida para estas comunidades virales termales. Esto podría estar representado por virus abundantes activos lisogénicos en El Tatio, que podrían infectar *Chloroflexota* (Figura 7, 8), lo que podría estar regulando el crecimiento y diversificación de este filo con cada vez más relevancia en el ecosistema termal a medida que aumenta la temperatura (Alcamán-Arias et al., 2018). Sin embargo, los productores primarios en estos ambientes se enfrentan a varias limitaciones en su crecimiento como son la solubilidad de los gases y degradación de clorofila a altas temperaturas, volviéndose altamente competitivas y generando una gran microdiversidad mediante mecanismos evolutivos antagonistas (KTW), producto de la gran actividad lítica (Guajardo- Leiva et al., 2021). Por tanto, nuestros resultados se condicen con la literatura, al ser los principales vOTUs los más activos (Figura 8), indicando que el principal rol que estarían cumpliendo estos virus sería infectar a los productores primarios más activos de estos tapetes microbianos, lisándolos acorde a la hipótesis KTW y controlando de esta manera su crecimiento y diversidad, además del ciclaje de nutrientes. Esta podría ser una de las explicaciones para la amplia distribución y diversidad de cianobacterias (Alcorta et al., 2020) y miembros del filo Chloroflexota en tapetes microbianos fototróficos termales.

7.5 Taxonomía de las comunidades virales de ambientes termales alrededor del mundo.

La secuenciación de virus y expansión de bases de datos genómicas virales han conducido al ICTV al consenso de cambiar el criterio de clasificación tradicional

(morfología del virión y filogenias de genes) hacia uno centrado en el genoma (Jang et al., 2019). Para bacteriófagos, un acercamiento temprano fue hacer comparaciones de secuencias de proteínas por pares de genomas completos en un marco filogenético (árbol proteómico de fagos), lo que fue ampliamente concordante con las agrupaciones virales avaladas por el ICTV en su momento (Rohwer y Edwards., 2002). Sin embargo, esta aproximación no fue extensamente adoptada al pensarse que el desenfrenado mosaicismo podría difuminar los límites taxonómicos y romper los supuestos de los algoritmos filogenéticos tras estos análisis (Lawrence et al., 2002). Otros acercamientos han estimado la fracción de genes compartidos y los cortes en el porcentaje de identidad para definir afiliaciones de género y subfamilias (Lavigne et al., 2009), no obstante, falló en definir la clasificación taxonómica para muchos virus conocidos, debido a la probabilidad de que la forma y tiempo en que los virus procariontes evolucionan son altamente variables (Mavrich y Hatfull., 2017). Finalmente, las redes de genes compartidos basadas en clústeres de proteínas (PCs) entre genomas virales han demostrado ser ampliamente concordantes con la taxonomía del ICTV independientemente del uso de redes monopartitas (genoma viral) o bipartitas (genoma viral más genes), siendo aplicadas para taxonomía viral de estudios a gran escala de océanos, aguas dulces y suelos, donde los virus solamente pudieron ser clasificados al aplicarse el método de red de genes compartidos (Jang et al., 2019).

Hasta el 2018, los fagos termales se encontraban taxonómicamente distribuidos en cincos familias: *Myoviridae*, *Siphoviridae*, *Inoviridae*, *Sphaerolipoviridae* y

Tectiviridae (Zablocki et al., 2018). Esto se basaba en una taxonomía acorde a la morfología de los virus al aislarlos en un número limitado de hospederos bacterianos termófilos, siendo *Thermus* el género modelo (Cava et al., 2009). Estas asignaciones son el resultado de investigaciones globales que se enfocaron en ambientes hipertermofílicos (72-98°C), debido al atractivo de aislar nuevos virus de arqueas, siendo las termas moderadamente termofílicas (40-71°C) pasadas por alto, pese a que probablemente son las que contienen la mayor abundancia y diversidad de termófagos autóctonos (Menzel et al., 2015). De esta manera, la revolución metaómica y consiguiente reestructuración de la clasificación viral, hacen de este estudio el primero en reclasificar las comunidades virales de bacteriófagos termales a escala regional y continental.

Globalmente, la gran mayoría de los virus identificados corresponden a la clase *Caudoviricetes* (Figura 11). Del total de vOTUs (3559) obtenidos en 49 muestras, un 40,89% logró ser asignado taxonómicamente. Mientras, del total de vOTUs (1844) clusterizados de las muestras, un 43.6% corresponden a secuencias que solamente se agrupan (clusterizan) entre sí, indicando una alta novedad taxonómica, al menos hasta el nivel de género viral (Roux et al., 2016, 2017, 2019; Martinez-Hernandez et al., 2017). Además, las abundancias relativas de los vOTUs en cada muestra indican que los virus más abundantes en prácticamente todas las muestras seleccionadas son desconocidos, denotando no solamente una alta novedad taxonómica individualmente a nivel de vOTU, sino también la importancia biológica de dichos virus por su alta abundancia (Gainer et al., 2017).

La clase más abundante identificada, *Caudoviricetes*, representa a los bacteriófagos y virus con cola que infectan arqueas (Evseev et al., 2023). Sin embargo, el rango de temperatura (40-80°C) y pH aproximadamente neutral (6-9) que se utilizó en este estudio (Figura 10), excluye a la mayoría de los procariontes acidófilos y arqueas hipertermofílicas (Salgado et al., 2021), dejando principalmente bacterias y arqueas termófilas que sería infectadas por bacteriófagos. Estos fagos juegan papeles muy importantes en la regulación y evolución de las comunidades microbianas presentes en los sistemas termales (Zablocki et al., 2018). Estos ambientes sirven cómo sistemas modelo para comprender la composición y función de estos consorcios microbianos (Guajardo-Leiva et al., 2018), al estar albergando comunidades microbianas simplificadas dominadas por una limitada variedad de organismos, siendo ideales para comprender las interacciones bióticas y cómo se moldea su estructura (Guajardo-Leiva et al., 2021). De igual manera, esta baja diversidad de hospederos puede también facilitar la caracterización de los virus presentes, observándose que a medida que se avanza en rangos taxonómicos, los órdenes virales más abundantes corresponden a los *Thumleimavirales*, *Crassvirales* y *Ligamenvirales* (Figura 12A). Junto a esto, se observan patrones geográficos de distribución taxonómica, evidenciándose que las comunidades virales son distintas según su localidad, a pesar de presentar condiciones fisicoquímicas similares, que se traduciría en similares hospederos a nivel de filo (Alcamán-Arias et al., 2018; Alcorta et al., 2020). A nivel de familia viral, también hay patrones biogeográficos, donde las principales familias virales (*Hafunaviridae*,

Herelleviridae, *Lavidaviridae* y *Rudiviridae*) se encuentran agrupadas según localidad termal específica (Figura 12b). Sin embargo, debido al bajo número de asignaciones taxonómicas respecto del total de abundancia viral presente en estos ambientes termales (Figura 11), estos resultados no pueden determinar con robustez los patrones biogeográficos por no ser representativos del total de la comunidad viral presente. Por ejemplo, en la figura 12b, la familia *Lavidaviridae* representa virófagos con altas abundancias para muestras de Huinay, Chile. Sin embargo, al considerar la totalidad de secuencias presentes en estas muestras, la clase que representa a dichos virus (Maveviricetes) corresponde a una abundancia casi imperceptible respecto a los Caudoviricetes y el resto de las secuencias virales desconocidas. Para superar esta limitación dada por la baja tasa de asignación taxonómica, se recurrió a un reclutamiento de secuencias contra todas las muestras utilizadas y posterior clusterización jerárquica junto a un cálculo de abundancias (Figura suplementaria 6). Este análisis evidenció que hasta el nivel de género (Roux et al., 2016; Martinez-Hernandez et al., 2017; Roux et al., 2017; Roux et al., 2019), existen patrones biogeográficos que indican una dispersión global limitada, mientras que una circulación y adaptación más local, a nivel termal y regional.

Así, se pone a prueba la hipótesis de Louren Baas Becking “Todo está en todas partes, pero el ambiente selecciona”, donde la discontinuidad ambiental que representa una terma, efectivamente funciona como un filtro ambiental (O’Malley., 2008), determinándose patrones biogeográficos locales en este caso, que podrían sugerir un potencial endemismo, así como también especiación

alopátrica (Papke et al., 2003; Menzel et al., 2015), al obtenerse un alto número de vOTUs exclusivos y abundantes de muestras que representan una misma localidad (Figura suplementaria 6). Esto puede ejemplificarse con las termas de El Tatio en Chile.

Es necesario destacar el hecho de que al momento de realizar asignaciones taxonómicas para estas secuencias virales y determinar diferencias entre las distintas regiones según su taxonomía, la nueva versión de la base de datos del IMG/VR aún se encuentra en constante remodelación. Es decir, dentro de la clase *Caudoviricetes* por ejemplo, hay 14 familias divididas en 4 órdenes virales, aunque existen 33 familias adicionales que están establecidas pero que aún no se han asignado a ningún orden viral (Abd-El Wahab et al., 2023). Un ejemplo puntual de nuestra investigación corresponde a la familia *Herelleviridae* (Figura 12b), una importante familia de bacteriófagos previamente descrita como *Siphoviridae* (Barylski et al., 2020), que tienen una alta representación en estos sistemas termales (Guajardo-Leiva et al., 2021), pero no se considera al momento de realizar asignaciones a niveles taxonómicos superiores (Orden).

Esta nueva resolución del ICTV y la revolución metaómica, están remodelando nuestra comprensión de la virosfera global (Camargo et al., 2022), implicando que los virus de este ambiente termal, en gran parte inexplorado, representa una caja negra aún más grande de lo esperado, provocando un renovado interés a través de la ventana de las ramas más profundas de la evolución y diversidad de los virus (Zablocki et al., 2018).

7.6 Diversidad de las comunidades virales de ambientes termales a lo largo de gradientes a escala global.

Los vOTUs capturan poblaciones evolutivas y ecológicamente cohesivas de genomas virales estrechamente relacionados (genotipos), las cuales no presentan diferencias de aptitud en el mismo espacio de nicho (hospedero) (Duhaime et al., 2017). Por lo tanto, los vOTUs proporcionan una unidad métrica para analizar comunidades virales a nivel de genoma y poblaciones. Modelos lineales generalizados (GLMs) mostraron en este estudio global, una significativa correlación positiva para la altitud ($p=0,019$), mientras que una correlación negativa para la latitud ($p=0,006$) y longitud ($p=0,028$), según la riqueza observada de especies.

La riqueza observada (para este caso, también índice Chao1) es un estimador de riqueza de especies que se basa en el número de clases (vOTUs) raras encontradas en una muestra, dándole más peso a estas especies (Borcard et al., 2018). Así, esta medida de diversidad alfa de especies aumenta en respuesta al aumento de altitud (posiblemente debido a la presencia en este estudio de un número elevado de termas de El Tatio localizadas a 4300 m sobre el nivel del mar), mientras que disminuye al recorrer un gradiente latitudinal de sur a norte y longitudinal de este a oeste (Figura 13a y 13b).

Los gradientes latitudinales de diversidad se caracterizan por una diversidad polar relativamente baja y alta en la zona ecuatorial, para la mayoría de la flora y fauna terrestre, así como también para el plancton oceánico (Dominguez-Huerta et al., 2022). Sin embargo, paradójicamente, la diversidad de los virus

procariontes de ADN de doble hebra marinos tiende a aumentar en el ártico, a diferencia de la diversidad de sus hospederos (Gregory et al., 2019). De igual manera, transectos en el océano atlántico han detectado patrones latitudinales de diversidad para virus gigantes que encuentran su máximo en las altas latitudes septentrionales, mientras que se estabilizan hacia el sur (A.D. Ha et al., 2023). En el caso de los sistemas termales, la discretización de las latitudes según su ubicación en los paralelos indica que, contrario a lo exhibido por los sistemas marinos, la diversidad de especies virales aumenta significativamente ($p=0,00008$) en el sur, mientras que en el hemisferio norte se mantiene en un margen reducido (Figura 13a). Aunque, al momento de establecer una distribución latitudinal equitativa entre ambos hemisferios, nos encontramos que la mayoría de las muestras utilizadas corresponden a latitudes bajas o medias (para ambos hemisferios) (Figura 10), siendo las muestras de bajas latitudes fueron significativamente ($p=0,01$) más diversas que las de latitudes medias (Figura suplementaria 3). Esto es concordante con el gradiente latitudinal de biodiversidad, donde el trópico es referido como una cuna de diversidad con alta especiación que se disipa a medida que avanza hacia los polos (Dowle et al., 2013). Trabajos de Mackenzie (2014) en tapetes microbianos termales en un gradiente global latitudinal a través de un transecto desde centro américa hasta la antártica reportaron que no había relaciones significativas entre la riqueza de especies bacterianas con la latitud, solamente al separar las muestras por temperatura (<50°C) encontraron relaciones significativas entre la riqueza observada de OTUs bacterianos y la latitud (indicando una disminución de la

diversidad al acercarse al meridiano de Greenwich desde el polo antártico) (Mackenzie, 2014)). Esta desconexión entre la diversidad de los virus y sus hospederos tiene precedentes entre simbiontes y sus hospederos eucariontes (Ibarbalz et al., 2019). Este fenómeno ha encontrado explicaciones en ambientes marinos, donde se ha hipotetizado que esta desconexión es causada por los diferentes impactos que la temperatura tiene en los virus y su hospedero, y/o que más virus de distintas especies puedan interactuar con el mismo hospedero (siendo la última hipótesis aún no testeada) (Dominguez- Huerta et al., 2022). Esta alusión a la promiscuidad de los virus podría ser una opción plausible en los ambientes termales, al ser dominado por pocos tipos de microorganismos y ser usualmente menos diversos que los sistemas acuáticos de bajas temperaturas y terrestres (Inskeep et al., 2010, 2013), explicando la posibilidad de una mayor diversidad viral en conjunto a una menor diversidad de hospederos.

La altitud expone diferencias significativas de diversidad alfa para la riqueza observada, representando una mayor diversidad viral a mayor altitud (Figura 13d). Previamente se asumía que la riqueza de especies a través de un gradiente altitudinal incrementaba universalmente desde tierras altas frías hacia tierras bajas cálidas, replicando el aumento latitudinal de riqueza desde latitudes frías a cálidas. Sin embargo, desde la aceptación del gradiente altitudinal como plantilla modelo para probar hipótesis de patrones a larga escala de diversidad, poco consenso se ha logrado (Nogués-Bravo et al., 2008). A nivel de bacteriófagos, se ha evaluado el efecto de la altitud en la actividad de estas

comunidades en suelos de los Alpes suizos (gradiente de 400 metros de altitud), encontrándose que los fagos se mantienen constantemente activos en los suelos a lo largo de todo el gradiente, mientras que la actividad bacteriana fuertemente declinaba al aumentar la altitud (Merges et al., 2021). Estudios de suelo en el este del Himalaya también mostraron la diversidad de cianobacterias en un rango de altitud entre 300 metros sobre el nivel del mar y 3500, revelando una fuerte disminución de especies con el aumento de la altitud (Choudhary y Singh., 2013). Para tratar de resolver la posible ambigüedad de interpretaciones respecto al gradiente altitudinal para nuestro set muestral de trabajo, discretizaciones que clasificaron el gradiente en baja, media y alta altitud, muestran que efectivamente las más altas altitudes se correlacionan positivamente de manera significativa con la riqueza observada (Figura 13d). siguiéndole la altitud baja y finalmente la media en riqueza de especies (sin diferencias significativas entre ellas) (Figura 13d). Estudios en biopelículas de arroyos han determinado que la composición y diversidad de la comunidad viral puede ser explicada determinísticamente por los cambios en la comunidad bacteriana y que dicha comunidad subyace al acoplamiento virus-bacteria, demostrado en un gradiente altitudinal (Bekliz et al., 2022). Sin embargo, los antecedentes mencionados en suelos respecto de los efectos contrastantes de la altitud en la actividad microbiana (Merges et al., 2021), las diversas condiciones climáticas, nutrientes, radiación, entre otras cosas, que se pueden dar a lo largo de un gradiente altitudinal (Choudhary y Singh., 2013) desafían esta dependencia del huésped para la determinación de la diversidad viral.

Además, estudios en sistemas termales de la meseta del Tíbet evidencian que la comunidad bacteriana en grandes altitudes es similar a la de bajas altitudes, sugiriendo que la elevación no sería un factor determinante de la distribución bacteriana global en campos geotérmicos (Huang et al., 2011). Estos resultados denotan la necesidad urgente de generar más conocimientos sobre la dinámica, estrategias de infección y consecuencias co-evolutivas de la depredación viral a lo largo de gradientes como lo es el altitudinal, evidenciando no solamente la falta de estudios virales a escala global en sistemas termales, sino también de su contraparte hospedera.

Otro gradiente de diversidad corresponde a los gradientes longitudinales, encontrándose, por ejemplo, que, en estudios de parásitos, no han sido tan bien documentados como los gradientes latitudinales, pero se han registrado centros de alta diversidad en comparación con otras áreas del mundo (Morand y Krasnov., 2010). En sistemas marinos, se han sugerido la presencia de barreras para la dispersión longitudinal, inhibiendo el movimiento de estas especies parasitarias a lo largo del planeta, siendo la más efectiva la barrera del pacífico oriental, la cual, debido a la falta de islas, reducen el potencial de dispersión de muchas especies (Morand y Krasnov., 2010). Además, la barrera del “nuevo mundo” representada por el continente americano, previene el movimiento longitudinal entre el pacífico oriental y el atlántico (Morris y Costello., 2017). En sistemas termales hay poco estudiado tanto en virus como hospederos.

En sistemas termales de China, Lau y Aitchison, han realizado estudios en transectos longitudinales que recorren hasta 380 kilómetros para caracterizar la

comunidad bacteriana en tapetes microbianos, sugiriendo la imperativa conservación de los sistemas termales, ya que las comunidades bacterianas de sistemas termales con similares características fisicoquímicas y sosteniendo niveles similares de riqueza de OTUs, son filogenéticamente distintas (Lau y Aitchison., 2009). No existiendo hasta la fecha para comunidades virales de sistemas termales ningún estudio en el gradiente longitudinal.

A nivel de comunidades virales, si se han descrito patrones longitudinales en biopelículas de arroyos alpinos, evidenciando una vez más, que los cambios en la composición de los virus puede ser debido a cambios en la comunidad bacteriana (Bekliz et al., 2022). Nuestros resultados con la discretización de la longitud según su ubicación de acuerdo con el meridiano de Greenwich (Figura 13b), refuerzan la idea de que existe una diferencia significativa en la diversidad viral según la ubicación en sistemas termales, habiendo mayor riqueza de especies en el hemisferio occidental, respecto del oriental. Por tanto, al evaluar en conjunto esta discretización de longitud junto a la de latitud (Figura 13a) y clusterización jerárquica de vOTUs (Figura suplementaria Heatmap), los datos sugieren la existencia de una biogeografía global de las comunidades virales de sistemas termales. Se presentan agrupaciones de secuencias virales distintas según su procedencia geográfica a nivel regional y/o continental (Figura suplementaria 6), y diferencias significativas en su diversidad según la riqueza observada (Figura suplementaria 4; Figura 13). Sin embargo, no podemos hablar en su totalidad de diferencias en la diversidad alfa para las comunidades virales en los electos sistemas termales, ya que, si bien la riqueza de especies

es la medida más simple de diversidad alfa, dicha diversidad debe estar compuesta tanto por la riqueza de especies como la equidad de estas (Thukral., 2017), no solamente por la riqueza (como en este caso, al no haber diferencias significativas para los otros índices utilizados).

Finalmente, no ha habido estudios sistemáticos que se hayan realizado para comparar las comunidades bacterianas del agua, tapete microbiano y sedimento de los sistemas termales (Chen et al., 2023), no existiendo así en el caso de las comunidades virales. De tal manera, Chen y colaboradores encontraron en sistemas geotermales de Taiwán que la mayor diversidad y abundancia de bacterias se encontró en el agua, seguido de los tapetes y finalmente el sedimento. Sin embargo, encontraron también una mayor abundancia de comunidades microbianas no identificadas en los tapetes microbianos y sedimentos, subrayando la necesidad de considerar estos tres hábitats como sistemas separados porque se ha observado que contienen distintos filos y géneros (Chen et al., 2023). En nuestro estudio global, contrario a lo encontrado para la contraparte bacteriana en el trabajo de Chen et al., 2023, la menor diversidad viral se encontró en el agua, aumentando en el sedimento, pero siendo más elevada en las muestras de tapete microbiano (Figura 13c), aunque solamente fue significativa la diferencia entre las muestras de agua y tapete microbiano (p-value = 0,036). Esta discordancia de riqueza de especies virales, respecto de lo reportado por la literatura para su contraparte bacteriana en hábitats termales, puede deberse a las distintas características fisicoquímicas, donde la literatura se inclina por sistemas termales ácidos (como es el caso de

Chen et al., 2023), mientras que el set de muestras presentadas en este trabajo corresponde a un pH neutro, el cuál a la temperatura seleccionada en este trabajo (40-80°C) promueve un gran crecimiento de distintos filos de bacterias en los tapetes microbianos (Alcamán et al., 2015).

Además, trabajos recientes de nuestro grupo de investigación en comunidades bacterianas de sistemas termales alrededor del mundo encontraron diferencias significativas para la diversidad de Shannon al agrupar las comunidades según hábitat, evidenciando la menor diversidad en las comunidades de agua, al compararlas con los tapetes microbianos (Barbosa et al., 2023). De esta forma se podría hipotetizar que, bajo estas condiciones, se puede estar dando una mayor diversidad de especies virales en los tapetes microbianos por sobre el agua y sedimento. Otro factor para considerar es el potencial sesgo para este trabajo en la cantidad de muestras que representan a cada uno de los hábitats, donde 31 de las 49 muestras corresponden a tapete, mientras que solamente 9 corresponden a sedimento y otras 9 a agua, no teniendo la misma robustez en su representatividad. Sin embargo, a diferencia de los tapetes termofílicos, la mayoría de la investigación viral en termas se ha enfocado en el agua de la fuente termal, encontrándose abundancias que van desde miles a millones de partículas del tipo viral por mililitro (Breitbart et al., 2004; Schoenfeld et al., 2008). Esta alta diversidad de especies de los tapetes microbianos también revela por tanto la alta novedad taxonómica que podría residir en este particular hábitat. Esto acentúa una vez más la necesidad de separar estos sistemas no sólo a nivel de comunidades bacterianas, sino que también virales, para así

poder identificar nuevos virus en sus respectivos patrones de distribución/coexistencia con sus potenciales hospederos en este gran ambiente que involucra distintos hábitats tan estrechamente relacionados.

A nivel de diversidad beta, se han hecho estudios de comunidades virales a escala regional y global en los océanos (Roux et al., 2016), donde se van estableciendo patrones biogeográficos que cada vez desafían más los modelos y gradientes de biodiversidad ya establecidos (Gregory et al., 2019).

En los 49 sistemas termales globales del presente trabajo, los factores ecológicos que determinan significativamente las diferencias de diversidad beta entre las comunidades virales son la latitud, longitud y altitud (forward selection, p-value-adj 0,010, 0,048 y 0,010, respectivamente), donde los análisis de redundancia evidencian patrones biogeográficos continentales de distribución de las comunidades virales (Figura 14), presentando comunidades similares entre distintos países que pertenecen al mismo continente, y generando agrupaciones independientes entre norteamérica, el sudeste asiático y latinoamérica (Figura 14). Esto indicaría que los virus efectivamente son sujetos de una limitada dispersión global, mientras que circulan de manera local, siendo la ubicación geográfica el principal determinante de la diferencia en la diversidad de especies virales, independientemente del hábitat del cuál provengan u otros factores fisicoquímicos (Figura 14; Tabla suplementaria 3).

En los sistemas geotermales, se ha descrito biogeografía microbiana a escala regional en muestras de agua de 925 termas de Nueva Zelanda, donde los patrones de decaimiento por distancia muestran una tendencia positiva con el

aumento de distancia, sugiriendo fuertemente la existencia de una limitación de dispersión entre campos geotermales individuales (Power et al., 2018). Esta Zona Volcánica de Taupo también ha llamado la atención en estudios regionales para estudiar la influencia de factores fisicoquímicos en la abundancia relativa, diversidad y prevalencia de bacterias y arqueas en sedimentos, encontrándose que en un rango de pH que oscila entre 2 y 7,5, junto a un rango de temperatura que va desde los 18°C hasta los 93°C, las bacterias y arqueas compartían abundancias similares en general para las comunidades microbianas termales (Sriaporn et al., 2023). En este estudio los autores identifican 8 variantes filogenéticamente diversas de arqueas y bacterias que se encontraron en el 84% de las comunidades termales, correspondiendo al 44% de la abundancia relativa a través de termas geográficamente distantes con amplios rangos de pH y temperatura, ilustrando la falta de restricción de la temperatura y el pH en su distribución (Sriaporn et al., 2023). A escala global, se han realizado análisis comparativos de diversidad microbiana a través de gradientes de temperatura entre una de las termas de Yellowstone (USA) y dos termas provenientes de Islandia, encontrándose potenciales linajes endémicos en arqueas, así como en bacterias de fila *Chloroflexota* y *Cyanobacteria* (Podar et al., 2020). Junto a esto, los trabajos en cianobacterias de Papke y colaboradores, que incluyen más ubicaciones y continentes, concluyen que las diferencias químicas no explican las diferencias en la distribución de especies a lo largo de los continentes, siendo así el aislamiento geográfico un importante factor para la divergencia evolutiva, además de ser un aspecto subestimado en la evolución microbiana (Papke et

al., 2003; Papke y Ward., 2004).

Sin embargo, a nivel de comunidades de bacteriófagos en ecosistemas termales, no existen estudios que describan patrones de diversidad beta y/o biogeografía, solamente revisiones taxonómicas de fagos provenientes de distintas localidades que han sido cultivados con los años (Zablocki et al., 2018). Los esfuerzos de dilucidar las comunidades virales termofílicas se han centrado principalmente en los virus que infectan arqueas, donde trabajos filogenómicos comparativos a escala global (Islandia, Estados Unidos, Portugal, Japón, Italia y México) han reportado fuertes patrones biogeográficos en la familia viral *Rudiviridae*, formando clados correspondientes a su origen geográfico, sugiriendo que la evolución y diversificación de los rudivirus estaría influenciada por el confinamiento espacial en estos sistemas hidrotermales continentales, con una pequeña migración horizontal de partículas virales sobre grandes distancias (Baquero et al., 2020).

De esta manera, al ser los virus entidades biológicas que dependen de su hospedero para replicarse y transcribirse, se podría hipotetizar que la biogeografía global determinada por este trabajo podría ir de la mano con la biogeografía de hospederos como en el caso de las cianobacterias (Papke y Ward, 2004). No obstante, esta biogeografía se ha descrito únicamente y hasta la fecha para las cianobacterias, las cuales representan a una minoría dentro del set global, no estando presente en varias de las termas investigadas en el presente trabajo (Salgado et al., en preparación). Junto a esto, parámetros como la temperatura y el pH que han demostrado tener la mayor contribución en

restringir a la diversidad microbiana (Hamilton et al., 2012; Menzel et al., 2015; Alcorta et al., 2018), no lograron superar el análisis de un “forward selection”, y presentaron el menor porcentaje de la varianza para las comunidades virales según un análisis multivariado permutado de la varianza (Tabla suplementaria 3). Lo anterior refleja la poca influencia que dichos factores podrían tener sobre la diversidad, mientras que a la vez reforzó a la geografía como la variable que más explica la distribución de la comunidad viral (Tabla suplementaria 3). Estos antecedentes podrían sugerir que la biogeografía de los bacteriófagos no depende directamente de la distribución global de sus hospederos, siendo plausible gracias a la posibilidad de que puedan infectar a distintos filos bacterianos, como sugiere el análisis del proteoma completo de fagos termales de distintas partes del mundo que ha mostrado clusterización entre fagos que infectan distintos filos hospederos (Zablocki et al., 2018).

Con todo esto, este trabajo muestra por primera vez una caracterización de las comunidades de bacteriófagos de campos geotermales en una escala espacial tanto local, regional y finalmente global, y los factores ecológicos que impulsan su diversificación, revelando por primera vez patrones biogeográficos para este tipo de comunidades virales en ambientes termales. Además, este trabajo representa un conjunto coherente de datos que puede ahora ser utilizado como base para futuros estudios de diversificación, deriva ecológica y que incluyan análisis más profundos de toda la comunidad microbiana, para dilucidar el efecto de distintos parámetros ambientales en la evolución microbiana en estos laboratorios naturales que son los campos termales.

8 Conclusiones

- Hay una alta novedad taxonómica en las comunidades virales que infectan bacterias en el campo geotermal El Tatio en Chile, al igual que en otros sistemas termales alrededor del mundo.
- La diversidad de especies virales en El Tatio está determinada principalmente por las coordenadas (latitud y longitud) y temperatura, existiendo una heterogeneidad espacial entre fuentes termales, mitigada por una dispersión local.
- Las comunidades virales más abundantes en los géiseres de El Tatio también son las comunidades más activas transcripcionalmente, siguiendo un modelo ecológico basado en la lisis viral.
- Los bacteriófagos del campo geotermal El Tatio infectan principalmente a los filos *Cyanobacteria*, *Chloroflexota*, *Bacteroidota* y *Proteobacteria*.
- Globalmente, patrones de diversidad de especies se correlacionan principalmente con las coordenadas geográficas, apoyando una biogeografía viral que desafía la hipótesis de Louren Baas Becking ("Todo está en todas partes, pero el ambiente selecciona").
- Se caracterizó por primera vez patrones biogeográficos en las comunidades de bacteriófagos de campos geotermales en una escala continental, evidenciando una limitada dispersión global, mientras que una circulación y adaptación local.
- Esta investigación propone al campo geotermal El Tatio como un sistema modelo ideal para poner a prueba los factores ecológicos que estructuran

las comunidades virales, proporcionando un conjunto de datos coherente que puede ser utilizado como una base para futuros estudios comparativos de virus en ecosistemas termales.

9 Referencias

Abd-El Wahab, A., Basiouni, S., El-Seedi, H. R., Ahmed, M. F. E., Bielke, L. R., Hargis, B., Tellez-Isaias, G., Eisenreich, W., Lehnherr, H., Kittler, S., Shehata, A. A., & Visscher, C. (2023). An overview of the use of bacteriophages in the poultry industry: Successes, challenges, and possibilities for overcoming breakdowns. In *Frontiers in Microbiology* (Vol. 14). Frontiers Media S.A.

Abdelkareem, A. O., Khalil, M. I., A Elbehery, A. H., & Elaraby, M. (2018). VirNet: Deep attention model for viral reads identification. *13th International Conference on Computer Engineering and Systems*, 623–626.

Abed, R. M. M., Kohls, K., Leloup, J., & De Beer, D. (2018). Abundance and diversity of aerobic heterotrophic microorganisms and their interaction with cyanobacteria in the oxic layer of an intertidal hypersaline cyanobacterial mat. *FEMS Microbiology Ecology*, *94*(2).

Alcamán, M. E., Alcorta, J., Bergman, B., Vásquez, M., Polz, M., & Díez, B. (2017). Physiological and gene expression responses to nitrogen regimes and temperatures in Mastigocladus sp. strain CHP1, a predominant thermotolerant cyanobacterium of hot springs. *Systematic and Applied Microbiology*, *40*(2), 102–113.

Alcamán, M. E., Fernandez, C., Delgado, A., Bergman, B., & Díez, B. (2015). The cyanobacterium Mastigocladus fulfills the nitrogen demand of a terrestrial hot spring microbial mat. *ISME Journal*, *9*(10), 2290–2303.

Alcamán-Arias, M. E., Pedrós-Alió, C., Tamames, J., Fernández, C., Pérez-Pantoja, D., Vásquez, M., & Díez, B. (2018). Diurnal changes in active carbon and nitrogen pathways along the temperature gradient in porcelana hot spring microbial mat. *Frontiers in Microbiology*, *9*(OCT).

Alcorta, J., Alarcón-Schumacher, T., Salgado, O., & Díez, B. (2020). Taxonomic Novelty and Distinctive Genomic Features of Hot Spring Cyanobacteria. *Frontiers in Genetics*, *11*.

Alcorta, J., Espinoza, S., Viver, T., Alcamán-Arias, M. E., Trefault, N., Rosselló-Móra, R., & Díez, B. (2018). Temperature modulates Fischerella thermalis ecotypes in Porcelana Hot Spring. *Systematic and Applied Microbiology*, *41*(6), 531–543.

Amgarten, D., Braga, L. P. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics*, *9*(AUG).

Andersen, K. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *BioRxiv*.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null Hypothesis Testing: Problems, Prevalence, and an Alternative. In *Source: The Journal of Wildlife Management* (Vol. 64, Issue 4).

Andersson, A. F., & Banfield, J. F. (2008). Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science*, *320*(5879), 1047–1050.

Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I., & Koonin, E. V. (2020). Seeker: Alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Research*, *48*(21), E121.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.

Baquero, D. P., Liu, Y., Wang, F., Egelman, E. H., Prangishvili, D., & Krupovic, M. (2020). Structure and assembly of archaeal viruses. In *Advances in Virus Research* (Vol. 108, pp. 127–164). Academic Press Inc.

Barbosa, C., Tamayo-Leiva, J., Alcorta, J., Salgado, O., Daniele, L., Morata, D., & Díez, B. (2023). Effects of hydrogeochemistry on the microbial ecology of terrestrial hot springs. *Microbiology Spectrum*, *11*(5).

Barylski, J., Kropinski, A. M., Alikhan, N. F., & Adriaenssens, E. M. (2020). ICTV virus taxonomy profile: Herelleviridae. *Journal of General Virology*, *101*(4), 362–363.

Bekliz, M., Pramateftaki, P., Battin, T. J., & Peter, H. (2022). Viral diversity is linked to bacterial community composition in alpine stream biofilms. *ISME Communications*, *2*(1).

Bennett, A. C., Murugapiran, S. K., & Hamilton, T. L. (2020). Temperature impacts community structure and function of phototrophic Chloroflexi and

Cyanobacteria in two alkaline hot springs in Yellowstone National Park. *Environmental Microbiology Reports*, *12*(5), 503–513.

Bennett, A. C., Murugapiran, S. K., Kees, E. D., Sauer, H. M., & Hamilton, T. L. (2022). Temperature and Geographic Location Impact the Distribution and Diversity of Photoautotrophic Gene Variants in Alkaline Yellowstone Hot Springs. *Microbiology Spectrum*, *10*(3).

Bhaya, D., Grossman, A. R., Steunou, A. S., Khuri, N., Cohan, F. M., Hamamura, N., Melendrez, M. C., Bateson, M. M., Ward, D. M., & Heidelberg, J. F. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME Journal*, *1*(8), 703–713.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., Turner, D., & Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology*, *37*(6), 632–639.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, *8*.

B. Bolduc and A. Zayed, py-VirMatcher, (2020), BitBucket repository, https://bitbucket.org/MAVERICLab/virmatcher/

Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F., & Young, M. (2012). Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-Dominated Yellowstone Hot Springs. *Journal of Virology*, *86*(10), 5562–5573.

Bolduc, B., Wirth, J. F., Mazurie, A., & Young, M. J. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. *ISME Journal*, *9*(10), 2162–2177.

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L., & Sullivan, M. B. (2017). IVirus: Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. In *ISME Journal* (Vol. 11, Issue 1, pp. 7–14). Nature Publishing Group.

Borcard, D., Gillet, F., & Legendre, P. (2018). *Numerical Ecology with R*. Springer International Publishing.

Breheny, P., & Burchett, W. (2013). *Visualization of Regression Models Using visreg*. http://CRAN.R-project.org/package=visreg

Breitbart, M. (2012). Marine viruses: Truth or dare. In *Annual Review of Marine Science* (Vol. 4, pp. 425–448).

Breitbart, M., Wegley, L., Leeds, S., Schoenfeld, T., & Rohwer, F. (2004). Phage Community Dynamics in Hot Springs. *Applied and Environmental Microbiology*, *70*(3), 1633–1640.

Brock, T. D., & Brock, M. L. (1966). Temperature optima for algal development in Yellowstone and Iceland hot springs. *Nature*, *209*(5024), 733–734.

Brum, J. R., Cesar Ignacio-Espinoza, J., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J. M., Gorsky, G., Gregory, A. C., Guidi, L., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Poulos, B. T., … Sullivan, M. B. (2015). Patterns and ecological drivers of ocean viral communities. *Science*, *348*(6237).

Camargo, A. P., Nayfach, S., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S. J., Reddy, T. B. K., Mukherjee, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Roux, S. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research*, *51*(D1), D733–D743.

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2023). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*.

Cava, F., Hidalgo, A., & Berenguer, J. (2009). Thermus thermophilus as biological model. In *Extremophiles* (Vol. 13, Issue 2, pp. 213–231).

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, *36*(6), 1925–1927.

Chen, J. S., Hussain, B., Tsai, H. C., Nagarajan, V., Koner, S., & Hsu, B. M. (2023). Analysis and interpretation of hot springs water, biofilms, and sediment bacterial community profiling and their metabolic potential in the

area of Taiwan geothermal ecosystem. *Science of the Total Environment*, *856*.

Chen, X., Weinbauer, M. G., Jiao, N., & Zhang, R. (2021). Revisiting marine lytic and lysogenic virus-host interactions: Kill-the-Winner and Piggyback-the-Winner. In *Science Bulletin* (Vol. 66, Issue 9, pp. 871–874). Elsevier B.V.

Craig Everroad, R., Otaki, H., Matsuura, K., & Haruta, S. (2012). Diversification of bacterial community composition along a temperature gradient at a thermal spring. *Microbes and Environments*, *27*(4), 374–381.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2).

Davison, M., Treangen, T. J., Koren, S., Pop, M., & Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS ONE*, *11*(9).

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. In *Nature Reviews Microbiology* (Vol. 18, Issue 3, pp. 125–138). Nature Research.

Dominguez-Huerta, G., Zayed, A. A., Wainaina, J. M., Guo, J., Tian, F., Pratama, A. A., Bolduc, B., Mohssen, M., Zablocki, O., Pelletier, E., Delage, E., Alberti, A., Aury, J.-M., Carradec, Q., Da Silva, C., Labadie, K., Poulain, J., Oceans, T., Bowler, C., … Sullivan, M. B. (2022). Diversity and ecological footprint of

Global Ocean RNA viruses. *Science*, *376*(6598), 1202–1208.

Dougherty, D., & Robbins, A. (1997). sed & awk: UNIX Power Tools. " O'Reilly Media, Inc.".

Dowle, E. J., Morgan-Richards, M., & Trewick, S. A. (2013). Molecular evolution and the latitudinal biodiversity gradient. In *Heredity* (Vol. 110, Issue 6, pp. 501–510).

Duhaime, M. B., Solonenko, N., Roux, S., Verberkmoes, N. C., Wichels, A., & Sullivan, M. B. (2017). Comparative omics and trait analyses of marine Pseudoalteromonas phages advance the phage OTU concept. *Frontiers in Microbiology*, *8*(JUL).

Eddy, S. R., & Durbin, R. (1994). RNA sequence analysis using covariance models. Nucleic acids research, 22(11), 2079-2088.

Evseev, P., Gutnik, D., Shneider, M., & Miroshnikov, K. (2023). Use of an Integrated Approach Involving AlphaFold Predictions for the Evolutionary Taxonomy of Duplodnaviria Viruses. *Biomolecules*, *13*(1).

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, *8*(6).

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., & Herskovits, A. A. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. In *Nature Reviews Microbiology* (Vol. 13, Issue 10, pp. 641–650). Nature Publishing Group.

Fichant, G. A., & Burks, C. (1991). Identifying Potential tRNA Genes in Genomic DNA Sequences. In *J. Mol. Biol* (Vol. 220).

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, *281*(5374), 237–240.

Fierer, N., & Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, *103*(3), 626–631.

Filée, J., Bapteste, E., Susko, E., & Krisch, H. M. (2006). A selective barrier to horizontal gene transfer in the T4-type bacteriophages that has preserved a core genome with the viral replication and structural genes. *Molecular Biology and Evolution*, *23*(9), 1688–1696.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.

Gainer, P. J., Pound, H. L., Larkin, A. A., LeCleir, G. R., DeBruyn, J. M., Zinser, E. R., Johnson, Z. I., & Wilhelm, S. W. (2017). Contrasting seasonal drivers of virus abundance and production in the North Pacific Ocean. *PLoS ONE*, *12*(9).

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.

George, C., Lim, C. X. Q., Tong, Y., & Pointing, S. B. (2023). Community structure

of thermophilic photosynthetic microbial mats and flocs at Sembawang Hot Spring, Singapore. *Frontiers in Microbiology*, *14*.

Gibson, Greg; Muse, Spencer V. (2009). *A Primer of Genome Science* (3rd ed.). Sinauer Associates. p. 84

Gregory, A. C., Zablocki, O., Zayed, A. A., Howell, A., Bolduc, B., & Sullivan, M. B. (2020). The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host and Microbe*, *28*(5), 724-740.e8.

Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., … Roux, S. (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, *177*(5), 1109-1123.e14.

Gregory, S. *Contig Assembly*. Encyclopedia of Life Sciences, 2005.

Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F., & Díez, B. (2018). Active crossfire between cyanobacteria and cyanophages in phototrophic mat communities within hot springs. *Frontiers in Microbiology*, *9*(SEP).

Guajardo-Leiva, S., Santos, F., Salgado, O., Regeard, C., Quillet, L., & Díez, B. (2021). Unveiling Ecological and Genetic Novelty within Lytic and Lysogenic Viral Communities of Hot Spring Phototrophic Microbial Mats. *Microbiology Spectrum*, *9*(3), e00694-21.

Gudbergsdóttir, S. R., Menzel, P., Krogh, A., Young, M., & Peng, X. (2016). Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. *Environmental Microbiology*, *18*(3), 863–874.

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, *9*(1).

Ha, A. D., Moniruzzaman, M., & Aylward, F. O. (2023). Assessing the biogeography of marine giant viruses in four oceanic transects. *ISME Communications*, *3*(1).

Hamilton, T. L., & Havig, J. (2023). Meet Me in the Middle: Median Temperatures Impact Cyanobacteria and Photoautotrophy in Eruptive Yellowstone Hot Springs. *MSystems*, *7*(1), e01450-21.

Hamilton, T. L., Vogl, K., Bryant, D. A., Boyd, E. S., & Peters, J. W. (2012). Environmental constraints defining the distribution, composition, and evolution of chlorophototrophs in thermal features of Yellowstone National Park. *Geobiology*, *10*(3), 236–249.

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T., & Bhaya, D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. *PLoS ONE*, *4*(1).

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., & Sullivan, M. B. (2017).

Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. In *ISME Journal* (Vol. 11, Issue 7, pp. 1511–1520). Nature Publishing Group.

Huang, Q., Dong, C. Z., Dong, R. M., Jiang, H., Wang, S., Wang, G., Fang, B., Ding, X., Niu, L., Li, X., Zhang, C., & Dong, H. (2011). Archaeal and bacterial diversity in hot springs on the Tibetan Plateau, China. *Extremophiles*, *15*(5), 549–563.

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 Bionformatics*, 1–11.

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., … Zinger, L. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179*(5), 1084-1097.e21.

Inskeep, W. P., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Rusch, D. B., Tringe, S. G., Macur, R. E., Jennings, R. de M., Boyd, E. S., Spear, J. R., & Roberto, F. F. (2013). Phylogenetic and functional analysis of metagenome sequence from high-temperature archaeal habitats demonstrate linkages between metabolic potential and geochemistry. *Frontiers in Microbiology*, *4*(MAY).

Inskeep, W. P., Rusch, D. B., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Richardson, T. H., Macur, R. E., Hamamura, N., Jennings, R. de M., Fouke, B. W., Reysenbach, A. L., Roberto, F., Young, M., Schwartz, A., Boyd, E. S.,

Badger, J. H., Mathur, E. J., Ortmann, A. C., Bateson, M., … Frazier, M. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS ONE*, *5*(3).

Jarett, J. K., Džunková, M., Schulz, F., Roux, S., Paez-Espino, D., Eloe- Fadrosh, E., Jungbluth, S. P., Ivanova, N., Spear, J. R., Carr, S. A., Trivedi, C. B., Corsetti, F. A., Johnson, H. A., Becraft, E., Kyrpides, N., Stepanauskas, R., & Woyke, T. (2020). Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME Journal*, *14*(10), 2527–2541.

Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., & Nielsen, M. (2016). MetaPhinder - Identifying bacteriophage sequences in metagenomic data sets. *PLoS ONE*, *11*(9).

Kamada, T., & Kawai, S. (1989). AN ALGORITHM FOR DRAWING GENERAL UNDIRECTED GRAPHS. *Information Processing Letters*, *31*(1), 7–15.

Kees, E. D., Murugapiran, S. K., Bennett, A. C., & Hamilton, T. L. (2022). Distribution and Genomic Variation of Thermophilic Cyanobacteria in Diverse Microbial Mats at the Upper Temperature Limits of Photosynthesis. *MSystems*, *7*(5).

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*, *8*(1).

Kishore Choudhary, K., & Singh, R. K. (2013). Cyanobacterial diversity along

altitudinal gradient in Eastern Himalayas of India. *Research Article J. Algal Biomass Utln*, *2*, 53–58.

Klatt, C. G., Inskeep, W. P., Herrgard, M. J., Jay, Z. J., Rusch, D. B., Tringe, S. G., Parenteau, M. N., Ward, D. M., Boomer, S. M., Bryant, D. A., & Miller, S. R. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. *Frontiers in Microbiology*, *4*(JUN).

Klatt, C. G., Wood, J. M., Rusch, D. B., Bateson, M. M., Hamamura, N., Heidelberg, J. F., Grossman, A. R., Bhaya, D., Cohan, F. M., Kühl, M., Bryant, D. A., & Ward, D. M. (2011). Community ecology of hot spring cyanobacterial mats: Predominant populations and their functional potential. *ISME Journal*, *5*(8), 1262–1278.

Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobian-Guëmes, A. G., Coutinho, F. H., Dinsdale, E. A., Felts, B., Furby, K. A., George, E. E., Green, K. T., Gregoracci, G. B., Haas, A. F., Haggerty, J. M., Hester, E. R., Hisakawa, N., Kelly, L. W., Lim, Y. W., … Rohwer, F. (2016). Lytic to temperate switching of viral communities. *Nature*, *531*(7595), 466–470.

Kohl, M., Wiese, S., & Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. Data mining in proteomics: from standards to applications, 291-303.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie

2. *Nature Methods*, *9*(4), 357–359.

Lau, M. C. Y., Aitchison, J. C., & Pointing, S. B. (2009). Bacterial community composition in thermophilic microbial mats from five hot springs in central Tibet. *Extremophiles*, *13*(1), 139–149.

Lavigne, R., Darius, P., Summer, E. J., Seto, D., Mahadevan, P., Nilsson, A. S., Ackermann, H. W., & Kropinski, A. M. (2009). Classification of myoviridae bacteriophages using protein sequence similarity. *BMC Microbiology*, *9*.

Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H. W., & Kropinski, A. M. (2008). Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Research in Microbiology*, *159*(5), 406–414.

Lawrence, J. G. (2002). Gene Transfer in Bacteria: Speciation without Species? *Theoretical Population Biology*, *61*(4), 449–460.

Lee, K. C. Y., Morgan, X. C., Dunfield, P. F., Tamas, I., McDonald, I. R., & Stott, M. B. (2014). Genomic analysis of Chthonomonas calidirosea, the first sequenced isolate of the phylum Armatimonadetes. *ISME Journal*, *8*(7), 1522–1533.

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research*, *46*(D1), D708-D717.

Liu, Y., Demina, T. A., Roux, S., Aiewsakun, P., Kazlauskas, D., Simmonds, P.,

Prangishvili, D., Oksanen, H. M., & Krupovic, M. (2021). Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. *PLoS Biology*, *19*(11).

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. In *Nucleic Acids Research* (Vol. 25, Issue 5).

Mackenzie Calderón, R. (2015). Ecology of hot spring microbial mats: Diversity, microheterogeneity, and biogeography. Universitat Autònoma de Barcelona.

Mackenzie, R., Pedrós-Alió, C., & Díez, B. (2013). Bacterial composition of microbial mats in hot springs in Northern Patagonia: Variations with seasons and temperature. *Extremophiles*, *17*(1), 123–136.

Maniloff, J., & Ackermann, H.-W. (1998). Virology Division News Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. In *VDN Virology Division News Arch Virol* (Vol. 143, Issue 10).

Marquet, M., Hölzer, M., Pletz, M. W., Viehweger, A., Makarewicz, O., Ehricht, R., & Brandt, C. (2022). What the Phage: a scalable workflow for the identification and analysis of phage sequences. *GigaScience*, *11*.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, *17*(1), 10–12.

Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., De La

Cruz Peña, M. J., Martínez, J. M., Anton, J., Gasol, J. M., Rosselli, R., Rodriguez-Valera, F., Sullivan, M. B., Acinas, S. G., & Martinez-Garcia, M. (2017). Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nature Communications*, *8*.

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A. L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. In *Nature Reviews Microbiology* (Vol. 4, Issue 2, pp. 102–112).

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*.

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, *8*(4).

Menzel, P., Gudbergsdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., Moracci, M., Kristjansson, J. K., Bolduc, B., Gavrilov, S., Ravin, N., Mardanov, A., Bonch-Osmolovskaya, E., Young, M., Krogh, A., & Peng, X. (2015). Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs. *Microbial Ecology*, *70*(2), 411–424.

Merges, D., Dal Grande, F., Greve, C., Otte, J., & Schmitt, I. (2021). Virus diversity in metagenomes of a lichen symbiosis (Umbilicaria phaea): complete viral genomes, putative hosts and elevational distributions.

Environmental Microbiology, *23*(11), 6637–6650.

Miller, S. R., Castenholz, R. W., & Pedersen, D. (2007). Phylogeography of the thermophilic cyanobacterium Mastigocladus laminosus. Applied and Environmental Microbiology, 73(15), 4751-4759.

Miller, S. R., Strong, A. L., Jones, K. L., & Ungerer, M. C. (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. *Applied and Environmental Microbiology*, *75*(13), 4565–4572.

Moon, K., & Cho, J. C. (2021). Metaviromics coupled with phage-host identification to open the viral 'black box.' In *Journal of Microbiology* (Vol. 59, Issue 3, pp. 311–323). The Korean Society for Mocrobiology / The Korean Society of Virology.

Morand, S., & Krasnov, B. R. (Eds.). (2010). The biogeography of host-parasite interactions. OUP Oxford.

Moreno, I. J., Brahamsha, B., Donia, M. S., & Palenik, B. (2023). Diverse Microbial Hot Spring Mat Communities at Black Canyon of the Colorado River. *Microbial Ecology*, *86*(3), 1534–1551.

Morris, T. C., & Costello, M. J. (2020). The biology, ecology and societal importance of marine parasites. In *Encyclopedia of the World's Biomes* (Vols. 4–5, pp. 556–566). Elsevier.

Munson-Mcgee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J.,

Weitz, J. S., & Young, M. J. (2018). A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments. *ISME Journal*, *12*(7), 1706–1714.

Muster, P., Binder, A., Schneider, K., & Bachofen, R. (1983). Influence of temperature and pH on the growth of the thermophilic cyanobacterium Mastigocladus laminosus in continuous culture. Plant and cell physiology, 24(2), 273-280.

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585.

New, F. N., & Brito, I. L. (2020). What Is Metagenomics Teaching Us, and What Is Missed? *Annual Review of Microbiology*, *74*, 117–135.

Nogués-Bravo, D., Araújo, M. B., Romdal, T., & Rahbek, C. (2008). Scale effects and human impact on the elevational species richness gradients. *Nature*, *453*(7192), 216–219.

Nübel, U., Bateson, M. M., Vandieken, V., Wieland, A., Kühl, M., & Ward, D. M. (2002). Microscopic examination of distribution and phenotypic properties of phylogenetically diverse Chloroflexaceae-related bacteria in hot spring microbial matst. *Applied and Environmental Microbiology*, *68*(9), 4593– 4603.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., & Wagner, H. (2022). vegan: Community Ecology Package. R package

version 2.5-7. 2020. Preprint at, 3-1.

O'Malley, M. A. (2008). "Everything is everywhere: but the environment selects": ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *39*(3), 314–325.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, *536*(7617), 425–430.

Papke, R. T., Ramsing, N. B., Bateson, M. M., & Ward, D. M. (2003). Geographical isolation in hot spring cyanobacteria. In *Environmental Microbiology* (Vol. 5, Issue 8).

Papke, R. T., & Ward, D. M. (2004). The importance of physical isolation to microbial diversification. In *FEMS Microbiology Ecology* (Vol. 48, Issue 3, pp. 293–303).

Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., & Ottonello, S. (1994). Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix anaylsis of transcriptional control regions. Nucleic acids research, 22(7), 1247-1256.

Pericard, P., Dufresne, Y., Couderc, L., Blanquart, S., & Touzet, H. (2018). MATAM: Reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. *Bioinformatics*, *34*(4), 585–591.

Podar, P. T., Yang, Z., Björnsdóttir, S. H., & Podar, M. (2020). Comparative Analysis of Microbial Diversity Across Temperature Gradients in Hot Springs From Yellowstone and Iceland. *Frontiers in Microbiology*, *11*.

Power, J. F., Carere, C. R., Lee, C. K., Wakerley, G. L. J., Evans, D. W., Button, M., White, D., Climo, M. D., Hinze, A. M., Morgan, X. C., McDonald, I. R., Cary, S. C., & Stott, M. B. (2018). Microbial biogeography of 925 geothermal springs in New Zealand. *Nature Communications*, *9*(1).

Prangishvili, D., & Krupovic, M. (2012). A new proposed taxon for double-stranded DNA viruses, the order “Ligamenvirales.” *Archives of Virology*, *157*(4), 791–795.

Pride, D. T., & Schoenfeld, T. (2008). Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics*, *9*.

Ramette, A. (2007). Multivariate analyses in microbial ecology. In *FEMS Microbiology Ecology* (Vol. 62, Issue 2, pp. 142–160).

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*(1), 69.

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*(1), 64–77.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor

package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rodriguez-Brito, B., Li, L. L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., Desnues, C., Dinsdale, E., Edwards, R., Felts, B., Haynes, M., Liu, H., Lipson, D., Mahaffy, J., Martin-Cuadrado, A. B., Mira, A., Nulton, J., Pašić, L., … Rohwer, F. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME Journal*, *4*(6), 739–751.

Rohwer, F., & Edwards, R. (2002). The phage proteomic tree: A genome-based taxonomy for phage. *Journal of Bacteriology*, *184*(16), 4529–4535.

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum information about an uncultivated virus genome (MIUVIG). *Nature Biotechnology*, *37*(1), 29–37.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., Poulos, B. T., Solonenko, N., Lara, E., Poulain, J., Pesant, S., Kandels-Lewis, S., Dimier, C., Picheral, M., Searson, S., Cruaud, C., Alberti, A., Duarte, C. M., Gasol, J. M., … Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, *537*(7622), 689–693.

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., & Sullivan, M. B. (2017).

Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, *2017*(9).

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *2015*(5).

Salgado, O., Guajardo-Leiva, S., Moya-Beltrán, A., Barbosa, C., Ridley, C., Tamayo-Leiva, J., Quatrini, R., Mojica, F. J. M., & Díez, B. (2022). Global phylogenomic novelty of the Cas1 gene from hot spring microbial communities. *Frontiers in Microbiology*, *13*.

Sano, D., Tazawa, M., Inaba, M., Kadoya, S., Watanabe, R., Miura, T., & Okabe, S. (2018). Selection of cellular genetic markers for the detection of infectious poliovirus. *Journal of applied microbiology*, *124*(4), 1001-1007.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., & Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Applied and Environmental Microbiology*, *74*(13), 4164–4174.

Schwabe GH (1960) Über den thermobionten kosmopolitan Mastigocladus laminosus Cohn. Blaualgen und Lebensraum V. Schweiz Z Hydrol 22:757–792

Sharma, A., Schmidt, M., Kiesel, B., Mahato, N. K., Cralle, L., Singh, Y., Richnow, H. H., Gilbert, J. A., Arnold, W., & Lal, R. (2018). Bacterial and Archaeal

Viruses of Himalayan Hot Springs at Manikaran Modulate Host Genomes. *Frontiers in Microbiology*, *9*.

Sharp, C. E., Smirnova, A. V., Graham, J. M., Stott, M. B., Khadka, R., Moore, T. R., Grasby, S. E., Strack, M., & Dunfield, P. F. (2014). Distribution and diversity of Verrucomicrobia methanotrophs in geothermal and acidic environments. *Environmental Microbiology*, *16*(6), 1867–1878.

Silveira, C. B., & Rohwer, F. L. (2016). Piggyback-The-Winner in host-Associated microbial Communities. In *npj Biofilms and Microbiomes* (Vol. 2). Nature Publishing Group.

Simmonds, P., Adriaenssens, E. M., Murilo Zerbini, F., Abrescia, N. G. A., Aiewsakun, P., Alfenas-Zerbini, P., Bao, Y., Barylski, J., Drosten, C., Duffy, S., Paul Duprex, W., Dutilh, B. E., Elena, S. F., García, M. L., Junglen, S., Katzourakis, A., Koonin, E. V., Krupovic, M., Kuhn, J. H., … Vasilakis, N. (2023). Four principles to establish a universal virus taxonomy. *PLoS Biology*, *21*(2).

Snyder, J. C., Wiedenheft, B., Lavin, M., Roberto, F. F., Spuhler, J., Ortmann, A. C., Douglas, T., & Young, M. (2007). Virus movement maintains local virus population diversity. *Proceedings of the National Academy of Sciences*, *104*(48), 19102–19107.

Sourrouille, Z. A., Schwarzer, S., Lequime, S., Oksanen, H. M., & Quax, T. E. F. (2022). The Viral Susceptibility of the Haloferax Species. *Viruses*, *14*(6).

Sriaporn, C., Campbell, K. A., Van Kranendonk, M. J., & Handley, K. M. (2023).

Bacterial and archaeal community distributions and cosmopolitanism across physicochemically diverse hot springs. *ISME Communications*, *3*(1).

Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N., & Govorun, V. M. (2020). Phigaro: High-throughput prophage sequence annotation. *Bioinformatics*, *36*(12), 3882–3884.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. In *Nature Biotechnology* (Vol. 35, Issue 11, pp. 1026–1028). Nature Publishing Group.

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Babin, M., Boss, E., Cochrane, G., de Vargas, C., Follows, M., Gorsky, G., Grimsley, N., Guidi, L., Hingamp, P., Iudicone, D., Jaillon, O., Kandels, S., Karp-Boss, L., Karsenti, E., Lescot, M., … Lombard, F. (2020). Tara Oceans: towards global ocean ecosystems biology. In *Nature Reviews Microbiology* (Vol. 18, Issue 8, pp. 428–445). Nature Research.

Suttle, C. A. (2007). Marine viruses - Major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(10), 801–812.

Thukral, A. K. (2017). A review on measurement of Alpha diversity in biology. *Agricultural Research Journal*, *54*(1), 1.

Titus Brown, C., & Irber, L. (2016). sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software*, *1*(5), 27.

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-

Zerbini, P., ... & Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, *168*(2), 74.

Uribe-Lorío, L., Brenes-Guillén, L., Hernández-Ascencio, W., Mora-Amador, R., González, G., Ramírez-Umaña, C. J., Díez, B., & Pedrós-Alió, C. (2019). The influence of temperature and pH on bacterial community composition of microbial mats in hot springs from Costa Rica. *MicrobiologyOpen*, *8*(10).

Van Hannen, E. J., Van Agterveld, M. P., Gons, H. J., & Laanbroek, H. J. (1998). Revealing genetic diversity of eukaryotic microorganisms in aquatic environments by denaturing gradient gel electrophoresis. *Journal of Phycology*, *34*(2), 206–213.

Vergara-Barros, P., Alcorta, J., Casanova-Katny, A., Nürnberg, D. J., & Díez, B. (2022). Compensatory Transcriptional Response of Fischerella thermalis to Thermal Damage of the Photosynthetic Electron Transfer Chain. *Molecules*, *27*(23).

Vidakovic, L., Singh, P. K., Hartmann, R., Nadell, C. D., & Drescher, K. (2017). Dynamic biofilm architecture confers individual and collective mechanisms of viral protection. *Nature Microbiology*, *3*(1), 26–31.

Ward, D. M. (2006). Microbial diversity in natural environments: Focusing on fundamental questions. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *90*(4), 309–324.

Ward, D. M., Ferris, M. J., Nold, S. C., & Bateson, M. M. (1998). A Natural View

of Microbial Biodiversity within Hot Spring Cyanobacterial Mat Communities. In *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS* (Vol. 62, Issue 4).

Whitaker, R. J., Grogan, D. W., & Taylor, J. W. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, *301*(5635), 976–978.

Wickham, H., & Chang, W. (2014). *Package "ggplot2" Type Package Title An implementation of the Grammar of Graphics*.

Zablocki, O., van Zyl, L. J., Kirby, B., & Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. *Viruses*, *9*(11).

Zablocki, O., van Zyl, L., & Trindade, M. (2018). Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages. In *Extremophiles* (Vol. 22, Issue 6, pp. 827–837). Springer Tokyo.

10 Anexos

Tabla Suplementaria 1.- Muestras (metagenomas) de sistemas termales de Chile y otras partes del mundo utilizados en esta tesis, con su respectiva metadata, para la realización de todos los análisis y estudios comparativos de sus comunidades virales. La tabla presenta la información de las muestras utilizadas, organizada en nueve columnas: Temperatura, pH, hábitat de donde se obtuvo el ADN, altitud, país de procedencia, continente, coordenadas U.T.M, latitud y longitud.

Hot Spring	*Temperature (°C)*	*pH*	*DNA Source*	*Altitude (msnm)*	*Country*	*Continent*	*UTM*	*Latitude*	*Longitude*
Kroner	44	6.1	mat	1	Antarctica	Antarctica	20E	-62.9409297	-60.5553751
Jinata Onsen 1	41.65	6.7	mat	29	Japan	Asia	54S	34.3179496	139.216024
Anhoni 1	43.5	7.5	water	374	India	Asia	44Q	22.65	78.36
Dewar Creek 1	44.5	8.15	sediment	1016	Canada	North America	11U	49.69558	-116.37452
Jinata Onsen 2	49.25	6.5	mat	29	Japan	Asia	54S	34.3179496	139.216024
San Vicente	54.5	6.7	water	2463	Colombia	South America	18N	4.8375	-75.53916
Shivlinga 1	46	8	mat	4392	India	Asia	44S	33.22955	78.35525
Shivlinga 2	46	8	mat	4392	India	Asia	44S	33.22955	78.35525
Cone Pool 1	45.6	8.1	mat	2140	USA	North America	11S	37.6906	-118.8444
Cone Pool 2	45.6	8.1	mat	2140	USA	North America	11S	37.6906	-118.8444
Huinay 1	48.6	6.9	mat	193	Chile	South America	18G	-42.45	-72.45999
Manikaran	50	7	mat	170	India	Asia	43S	32.3333	72.35
Anhoni 2	52.1	7.8	water	374	India	Asia	44Q	22.65	78.36
Anhoni 3	55	7.8	water	374	India	Asia	44Q	22.65	78.36
Tattapani 1	55	7.9	water	602	India	Asia	44Q	23.41	83.38999
Huinay 2	58	6.9	mat	193	Chile	South America	18G	-42.45	-72.45999
Mushroom 1	60	8	mat	2196	USA	North America	12T	44.555	-110.83499
Mushroom 2	59.9	8	mat	2196	USA	North America	12T	44.555	-110.83499
Tattapani 2	61.5	7.6	water	602	India	Asia	44Q	23.41	83.38999
Er-Yuan	65	7	sediment	2137	China	Asia	47R	26.25722	99.99027
Dewar Creek 2	64.7	7.94	sediment	1016	Canada	North America	11U	49.69558	-116.37452
Washburn	67.5	6.4	sediment	2503	USA	North America	12T	44.7660494	-110.429645
Huinay 3	66	6.9	mat	193	Chile	South America	18G	-42.45	-72.45999
Dewar Creek 3	66.4	7.9	sediment	1016	Canada	North America	11U	49.69558	-116.37452
Tattapani 3	67	7.8	water	602	India	Asia	44Q	23.41	83.38999
Kirishima	68	6.9	water	746	Japan	Asia	52R	31.91596	130.79907

Tattapani 4	69	7	water	602	India	Asia	44Q	23.41	83.38999
Miyagi	70	7.2	sediment	313	Japan	Asia	54S	38	140
Mammoth	72	6.5	mat	1923	USA	North America	12T	44.97287	-110.70441
Gongxiaoshe 1	71.7	7.5	sediment	1792	China	Asia	47R	25.44012	98.44081
Gongxiaoshe 2	73.8	7.29	sediment	1792	China	Asia	47R	25.44012	98.44081
Jinze	78.2	6.7	sediment	1122	China	Asia	47Q	23.44138	98.46003
Tatio 60	60	6.8	mat	4259	Chile	South America	19K	-22.3375799	-68.1723242
Tatio 3(45)	45	7.2	mat	4282	Chile	South America	19K	-22.3452913	-68.1230384
Tatio 11(48)	48	8.61	mat	4262	Chile	South America	19K	-22.3366128	-68.1863738
Tatio 11(62)	62	8.54	mat	4262	Chile	South America	19K	-22.3366128	-68.1863738
Tatio 1	45	7.3	mat	4261	Chile	South America	19K	-22.334854	-68.12976
Tatio 10	55	7.53	mat	4302	Chile	South America	19K	-22.3463061	-68.838344
Tatio 11	56	8.63	mat	4262	Chile	South America	19K	-22.3366128	-68.1863738
Tatio 12	57	8	mat	4260	Chile	South America	19K	-22.3371114	-68.1893489
Tatio 13	50	9.27	mat	4283	Chile	South America	19K	-22.3558333	-68.2279506
Tatio 2	55.4	7.83	mat	4259	Chile	South America	19K	-22.3375799	-68.1723242
Tatio 3	55	7.15	mat	4282	Chile	South America	19K	-22.3452913	-68.1230384
Tatio 4	55	7.83	mat	4278	Chile	South America	19K	-22.3558975	-68.229597
Tatio 5	55	7.3	mat	4259	Chile	South America	19K	-22.3371203	-68.1742012
Tatio 6	55	7.44	mat	4261	Chile	South America	19K	-22.3366596	-68.144425
Tatio 7	50	8.24	mat	4266	Chile	South America	19K	-22.3332894	-68.1281558
Tatio 8	55	7.58	mat	4293	Chile	South America	19K	-22.3474353	-68.1134682
Tatio 9	54.3	7.69	mat	4319	Chile	South America	19K	-22.3501071	-68.807494

Tabla Suplementaria 2.- Test de normalidad de Shapiro-Wilks en el grupo de datos correspondiente a El Tatio y el grupo global, para cada índice de diversidad utilizado. Se representa el valor estadístico W y su valor estadístico p (p-value) para evaluar la normalidad de ambos grupos de datos para cada índice de diversidad. Un p-value > 0.05 y W cercano a 1 indica normalidad de los datos.

	El Tatio		**Global**	
Index	*W*	*p-value*	*W*	*p-value*
Observed	0.92182	0.3341	0.94219	0.0259
Pielou	0.93043	0.4152	0.76542	0.0000004479
Shannon	0.94877	0.6282	0.94558	0.03462
Simpson	0.88527	0.1213	0.40093	2.594e-12
Shannon Evenness	0.96396	0.8133	0.96844	0.2097

Tabla Suplementaria 3. Análisis multivariado permutado de la varianza (PERMANOVA) para las variables ambientales disponibles para ambos grupos de datos (El Tatio y Global). Se representa el valor estadístico p de significancia (p-value) y el porcentaje de la varianza para cada variable. Se omite en el caso del Tatio las coordenadas U.T.M y hábitat, debido a que son iguales para todos los puntos utilizados.

	El Tatio		**Global**	
Variables	R^2	*p-value*	R^2	*p-value*
Temperature	0.17845	0.047	0.04067	0.017
pH	0.06612	0.566	0.04458	0.008
Altitude	0.10921	0.226	0.11693	0.001
Latitude	0.11291	0.207	0.11728	0.001
Longitude	0.18988	0.024	0.07511	0.001
UTM	--	--	0.52367	0.001
DNA Source	--	--	0.13553	0.001

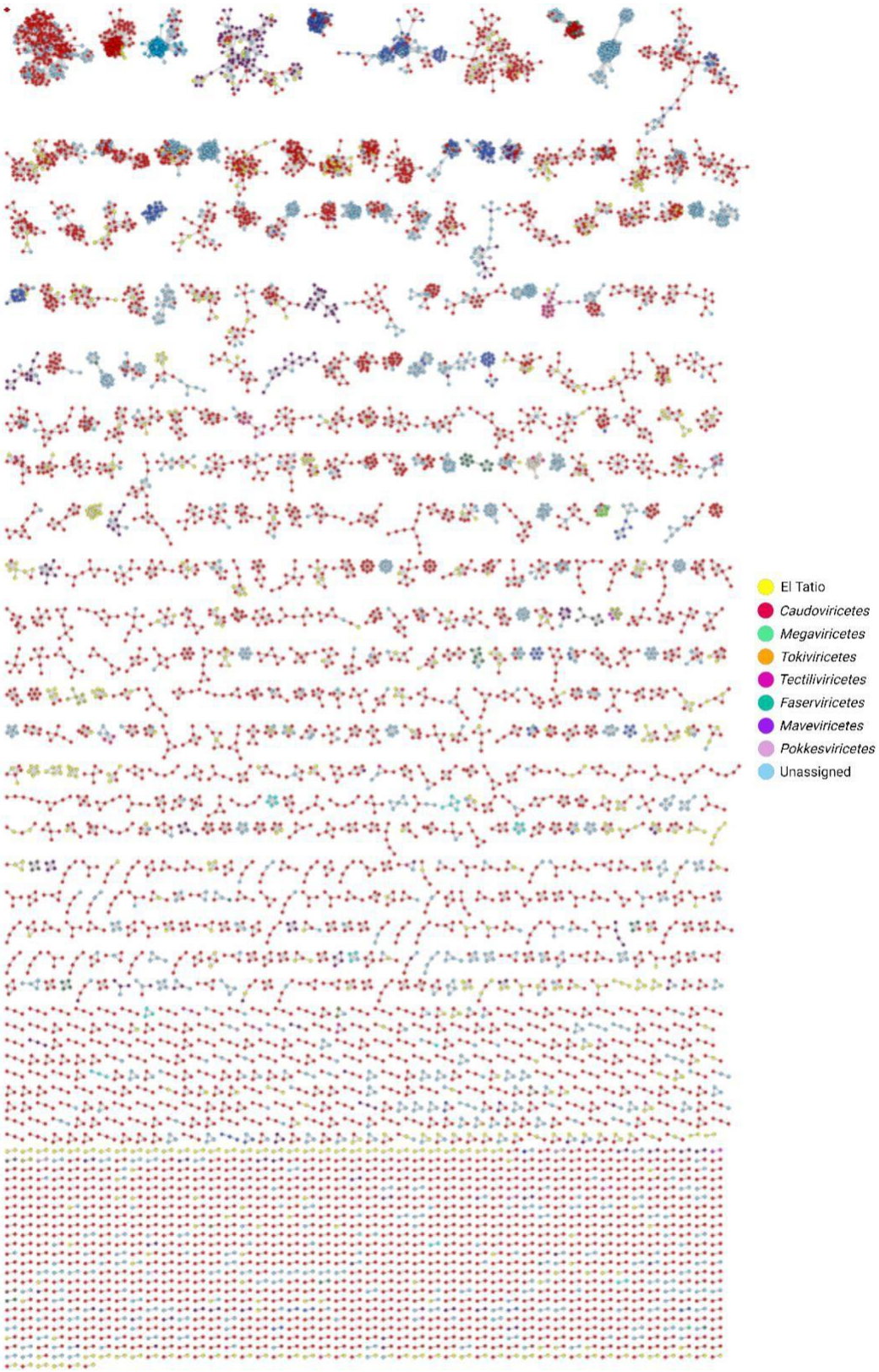
El Tatio
Caudoviricetes
Megaviricetes
Tokiviricetes
Tectiliviricetes
Faserviricetes
Maveviricetes
Pokkesviricetes
Unassigned

Figura Suplementaria 1 (página anterior).- Red completa de genes compartidos entre el set global de vOTUs (49 metagenomas) y la base de datos del IMG/VR v4 de sistemas termales. Red biológica completa obtenida con los resultados de vConTACT2 0.9.19 (Jang et al., 2019), y representados mediante Cytoscape 3.9.1 (Kohl et al., 2011) a través del algoritmo “Edge-Weighted Spring Embedded Layout” (Kamada y Kawai., 1988). Se representa la totalidad de los nodos que entran en la red (singletons suelen quedar fuera de la red) con la misma legenda utilizada previamente, observándose el total de clusters formados y las agrupaciones que siguen el resto de los virus de El Tatio entre sí y con las referencias del IMG/VR.

Figura Suplementaria 2.- Red de genes compartidos entre el set global de vOTUs (~50 metagenomas) y la base de datos del IMG/VR v3 de sistemas termales junto al RefSeq viral v201. Red biológica obtenida en análisis previos con vConTACT2 0.9.019 (Jang et al., 2019) y Cytoscape 3.8.2 (Kohl et al., 2011) con el algoritmo “Edge-Weighted Spring Embedded Layout” (Kamada y Kawai, 1998), que se realizaron para este trabajo de tesis, con las bases de datos disponibles a esa fecha (2022). Los nodos (genomas virales) están coloreados según sus familias virales (taxonomía antigua del ICTV) o si provienen de El Tatio.

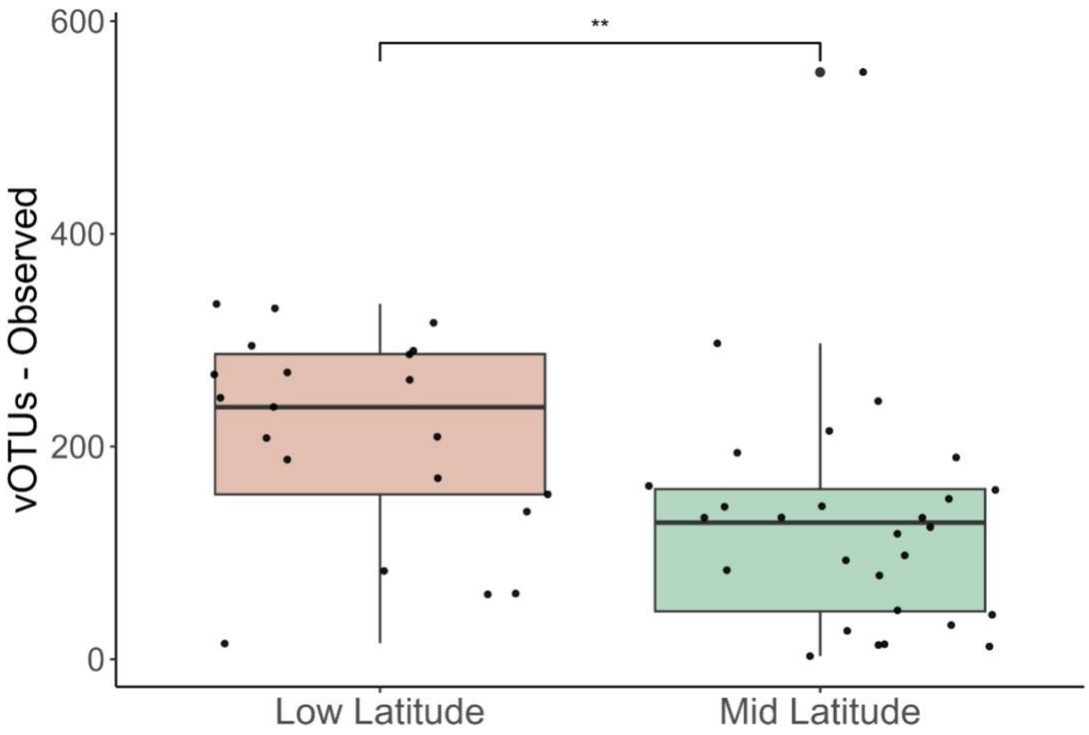

Figura Suplementaria 3.- Diagrama de cajas para los valores discretizados de latitud según su posición a lo largo de sus hemisferios respectivos, en 49 metagenomas globales, utilizando el índice de diversidad Riqueza observada (Observed). Se muestra un gráfico de cajas aplicado a la discretización de la latitud independientemente del hemisferio al que correspondan (las bajas latitudes cercanas al ecuador y altas latitudes (en este caso medias) cercanas a los polos), de acuerdo con la riqueza observada como índice de diversidad alfa, p-value = 0.01.

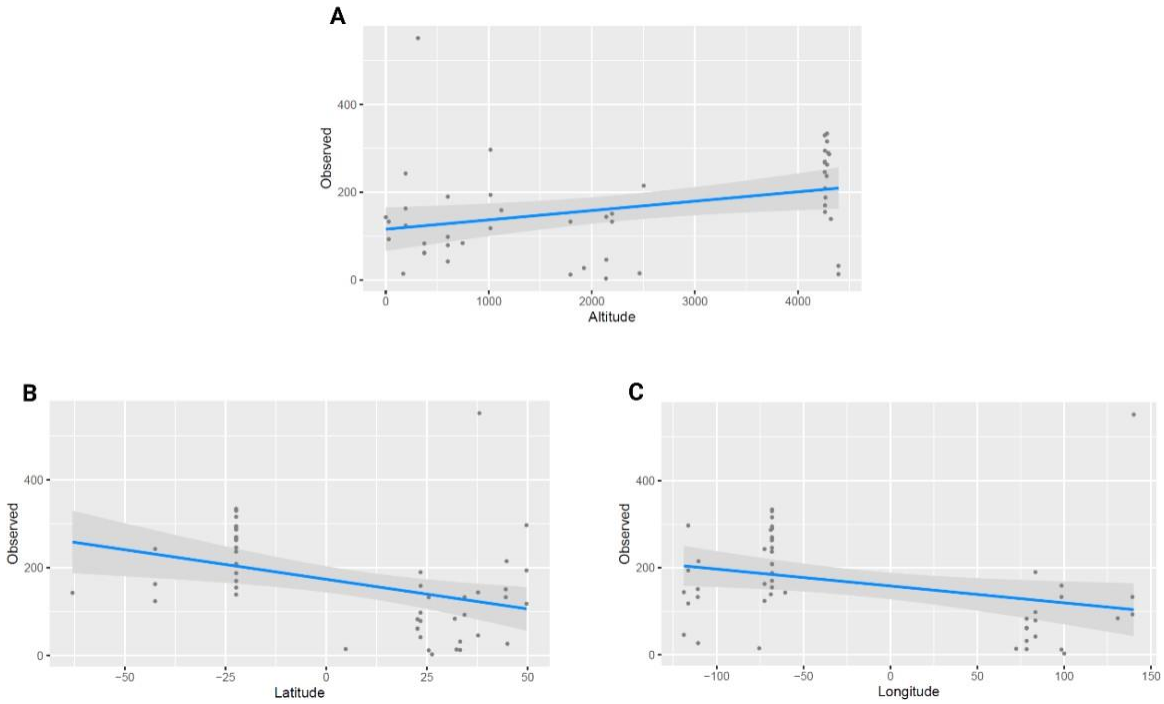

Figura Suplementaria 4.- Modelo Lineal Generalizado (GLM) simple en un gradiente de altitud, latitud y longitud para los 49 metagenomas del set global, utilizando el índice de diversidad alfa riqueza observada (Observed).

(A) GLM aplicado al gradiente altitudinal, de acuerdo con la riqueza observada como índice de diversidad alfa, obteniéndose un p-value de 0.01953.
(B) GLM aplicado al gradiente latitudinal utilizando el índice de riqueza observada, para representar el cambio de diversidad alfa, con un p-value de 0.0006549.
(C) GLM aplicado al gradiente longitudinal según la riqueza observada como métrica de diversidad alfa, teniendo un p-value de 0.0286.

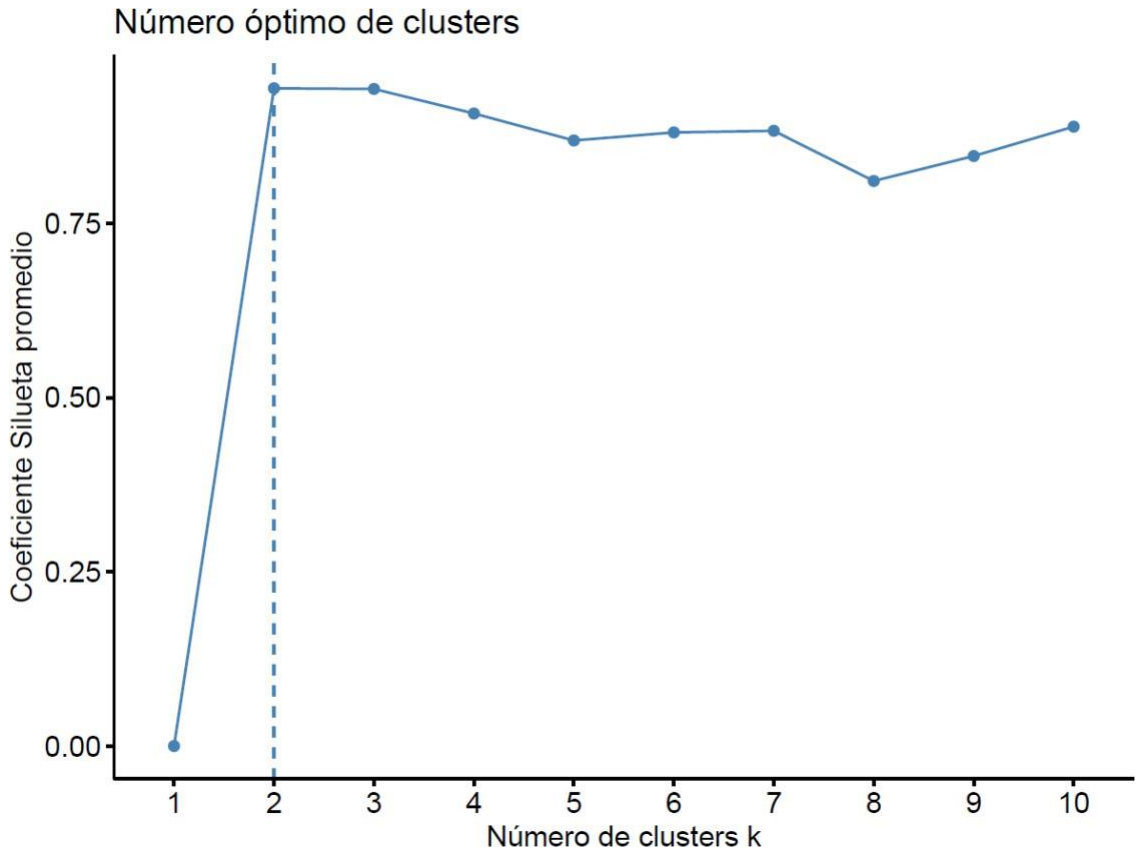

Figura Suplementaria 5.- Gráfico de siluetas para diferentes valores de k para el grupo correspondiente a los 49 metagenomas de sistemas termales alrededor del mundo. Se representa el cálculo de la media de los coeficientes de silueta de todas las muestras utilizadas (49) según sus distancias de bray-curtis, para establecer el número óptimo de clústers k para el set de datos, de forma que maximice la media de los coeficientes de silueta para dicho rango de clústers, al ser un valor alto indicativo de buen emparejamiento de grupos.

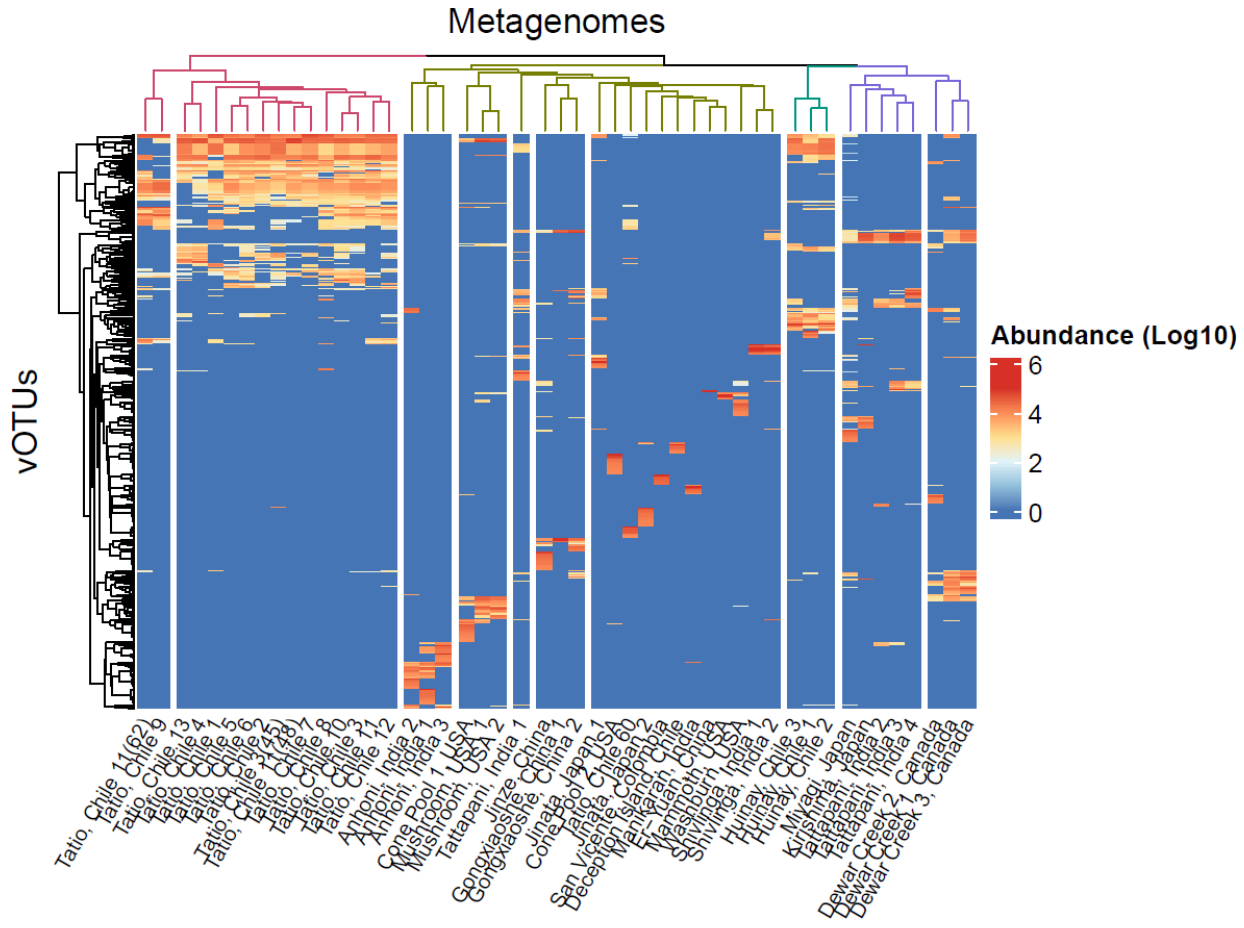

Figura Suplementaria 6.- Mapas de calor representando la abundancia absoluta de los vOTUs más abundantes (>1%) pertenecientes a 49 metagenomas y ordenados/agrupados según una clusterización jerárquica. A partir de los 3559 vOTUs obtenidos desde 49 metagenomas globales, se calcularon los vOTUs más abundantes (>1%), dando a lugar a 768 secuencias. En la figura se representan verticalmente los vOTUs más abundantes (>1%), visualizándose su abundancia en función de un gradiente de colores (mayor presencia atribuida al color rojo y menor al color azul) y su agrupación de acuerdo con un "clustering" jerárquico (dendrograma) según sus abundancias. La parte superior representa un clustering jerárquico en 4 grupos principales (k means = 4) para las muestras utilizadas, mientras que en la parte inferior se detalla la procedencia de cada muestra.