Pontificia Universidad Católica de Chile
Facultad de Ciencias Biológicas
Programa de Doctorado en Ciencias Biológicas
Mención Genética Molecular y Microbiología

TESIS DOCTORAL:

VIRAL COMMUNITIES OF PORCELANA HOT SPRINGS: CHARACTERIZATION AND ECOLOGICAL FUNCTION.

Por

SERGIO EDUARDO GUAJARDO LEIVA.

JUNIO 2019

Pontificia Universidad Católica de Chile
Facultad de Ciencias Biológicas
Programa de Doctorado en Ciencias Biológicas
Mención Genética Molecular y Microbiología

TESIS DOCTORAL:

VIRAL COMMUNITIES OF PORCELANA HOT SPRINGS: CHARACTERIZATION AND ECOLOGICAL FUNCTION.

Tesis presentada a la Pontificia Universidad Católica de Chile como parte de los requisitos para optar al grado de Doctor en Ciencias Biológicas mención Genética Molecular y Microbiología

Por

SERGIO EDUARDO GUAJARDO LEIVA.

Director de Tesis:	Dra. Beatriz Díez.
Comisión de Tesis:	Dr. Marcelo Cortez.
	Dr. Pablo González.
	Dr. Rodrigo De la Iglesia
	Dr. Rodrigo Gutierrez.

JUNIO 2019

“In the end we will conserve only what we love, we will love only what we understand, and we will understand only what we are taught”

Baba Dioum

DEDICATORIA

A Daniela por su amor e infinita paciencia,

a mi Madre por su dedicación

y a Matilde por existir...

AGRADECIMIENTOS

Solo espero no olvidar a nadie…

Primero a mi familia, la que construí y la que me vio nacer sin duda sin ellos nada de esto habría sido posible, muchas gracias por su apoyo incondicional.

A mi tutora Beatriz… quien me ayudo a sacar adelante esta tesis y me enseño a sobreponerme al fracaso. También por confiar en mi y siempre dejarme elegir, entregándome tu visión pero nunca imponiendo tu parecer sobre lo que yo pensé como correcto.

Bea, eres una excelente persona y científico, integra en ambos aspectos y siento que tuve suerte de elegirte como tutora y de que me hayas aceptado como estudiante.

A Eduardo por sus consejos y experiencias, pero sobre todo por su amistad.

A todo mi querido Lab BD los que pasaron y los que quedan, sin duda le han puesto sazón a este largo viaje Tomás, Jaime, Blanquis, Estrella, Jero, Pablo, Sebastian, Javier, Oscar, Octavio, Fabián, Ricardo, Cynthia, Cate, Brenda. También a los compañeros del Lab MV... Derly, Hector y Carol gracias por los cafecitos y la conversa. Gracias también por el babyshower mas épico que se realizado en el universo.

En fin, GRACIAS, GRACIAS y más GRACIAS!!!

A los que hablan y no callan, gracias
A los que quieren por lo que es y no por lo que vale, gracias
A los que lloran por que otros ríen y no ríen por que otros lloran, gracias
A los que van de frente y no por detrás, gracias
A los que sueñan y no duermen, gracias
A los que buscan problemas y no soluciones, gracias
A los que desordenan la vida y no se acomodan en ella, gracias
A los que se preguntan y no se responden, gracias
A los que cuestionan y no asienten, gracias
A los que me brindan seguridad con todas sus dudas, gracias
A los que creen en otra persona y no a otra persona, gracias
A los que creen en la búsqueda y no buscan a quien creerle, gracias
A los que creen en las causas y no en las causales, gracias
A los que creen en el sacrificio y no en sacrificar, gracias
A los que sospechan que no son libres, gracias
A los que saben que les falta algo y que ese algo no se compra, gracias
A los que resisten, a los que asisten, a los que dan pelea, gracias
Gracias por no recetar el remedio antes de encontrar la enfermedad
Y no inventar una infección para vendernos la cura
Gracias por tratar de atacar los motivos y no las consecuencias
Por enseñarnos que el saber no es inteligencia
Y que un libro no es sapiencia elitista sino herramienta popular
Gracias por interrogar e interrogarse y cuestionar la aglomeración de voluntades promoviendo la acción colectiva
Por demostrarnos que todos somos iguales en nuestras diferencias
Sin mejores ni peores pero con muchos diferentes
Que los opuestos se atraen y que los límites son barreras que nos bloquean
Gracias por pelear contra los prejuicios que a todos nos aquejan
Por reconocerlos y no negarlos
Para verlos, para tratar de derribarlos.

...Camino a Idilia

Este trabajo fue financiado por la beca de doctorado nacional CONICYT N° 21130667 y el proyecto FONDECYT N° 1150171.

INDEX

DEDICATORIA 3

AGRADECIMIENTOS 4

INDEX 6

ABBREVIATIONS 8

RESUMEN 10

ABSTRACT 14

GENERAL INTRODUCTION 18

HYPOTHESIS 30

GENERAL AIM 31

SPECIFIC AIMS 32

CHAPTER I 33

Active crossfire between Cyanobacteria and Cyanophages in phototrophic mat communities within hot springs

CHAPTER II 84

Killing the winner and piggybacking the cheater: lytic and lysogenic viral communities in hot springs phototrophic mats.

CHAPTER III 138

Ecological drivers modulate biogeography in thermophilic viral communities

GENERAL DISCUSSION **194**

GENERAL CONCLUSIONS **219**

REFERENCES **221**

PUBLICATIONS **231**

ABBREVIATIONS

YNP: Yellowstone National Park.

FAPs: Filamentous Anoxygenic Phototrophs.

Bchla: Bacteriochlorophyll A.

DOC: Dissolved Organic Carbon.

POC: Particulate Organic Carbon.

KtW: Kill the Winner.

PtW: Piggyback the Winner.

SVZ: Southern Volcanic Zone.

TEM: Transmission Electron Microscopy.

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats.

SNV: Single Nucleotide Variants.

MitC: mitomycin C.

MAG: Metagenome Assembled Genome.

GTAs: Gene Transfer Agents.

T6SSs: Type 6 Secretion Systems.

VLPs: Viral Like Particles.

vPCs: viral Protein Clusters.

vOTUs: viral Operational Taxonomic Units.

VCs: Viral Clusters.

GOS: Global Ocean Sampling.

POV: Pacific Ocean Virome.

TOV: Tara Ocean Virome.

Permanova: Permutational multivariate analysis of variance

PCoA: Principal Coordinates Analysis.

RESUMEN

Los virus han demostrado ser ubicuos en todos los ambientes por lo que las aguas termales no son la excepción a pesar de sus condiciones extremas. En aguas termales terrestres con pH circumneutral y temperaturas termofílicas, los tapetes fototróficos han demostrado ser útiles como modelos para comprender la composición, estructura y función de las comunidades microbianas en la naturaleza. Por lo tanto, proporcionan un marco teórico en el que estudiar el componente viral de estas comunidades y las interacciones virus-hospedero.

La Patagonia Norte es una de las regiones más australes del "Arco Volcánico Andino" y alberga 13 de los volcanes más activos de Chile. En consecuencia, una gran cantidad de fuentes termales se encuentran dispersas por los fiordos y bosques de la Patagonia. Algunos ejemplos son Porcelana y Cahuelmó, ubicados en los fiordos Comau y Cahuelmó, respectivamente. Ambas están cubiertas por tapetes microbianos que crecen a lo largo de un gradiente térmico entre 70 - 46°C, dominadas por fotótrofos bacterianos oxigénicos y anoxigénicos. El primer grupo está representado por los géneros *Fischerella*, *Calothrix*, *Leptolyngbya* y *Oscillatoria*, mientras que el grupo anoxigénico está representado por los géneros *Chloroflexus* y *Roseiflexus*. Otros taxones importantes en estas comunidades microbianas son las bacterias heterótrofas de los phyla Proteobacteria, Bacteroidetes y Deinococcus termus.

En la presente tesis, la recuperación de secuencias virales desde metagenomas y metatranscriptomas celulares de Porcelana, mostró que la comunidad viral está compuesta predominantemente por Caudovirales (70%), con la mayoría de las infecciones transcripcionalmente activas causadas por cianófagos (hasta el 90% de los transcritos de Caudovirales). El ensamblaje metagenómico permitió recuperar y describir el primer cianófago termofílico tipo T7 (TC-CHP58). Además, hemos encontrado marcadas diferencias en el número de loci CRISPR metagenómicos y en la diversidad de espaciadores (distintas secuencias y distintas abundancias) en *Fischerella,* así como Variantes de Nucleótido Único (SNV), en los proto-espaciadores de TC-CHP58 a diferentes temperaturas, lo que refuerza la teoría de la coevolución entre las poblaciones de virus naturales y sus cianobacterias hospederas.

Más tarde, estudiamos las comunidades virales líticas y lisogénicas en los tapetes fototróficos de Porcelana utilizando un enfoque multi-ómico junto a inducciones *in-situ* usando mitomicina C. Para esto, los genomas ensamblados de metagenomas (MAGs) de los tapetes microbianos de porcelana fueron interrogados sobre la presencia de virus integrados (Profagos). Así mismo, se analizaron los metagenomas de las comunidades virales naturales e inducidas por mitomicina C, para estudiar la abundancia diferencial de genomas virales y proteínas (funciones) entre ambas comunidades virales, así como parámetros ecológicos tales como las diversidades α y β. Nuestros resultados sugieren que las poblaciones virales lisogénicas y líticas estaban altamente asociadas con hospederos específicos. La mayoría de los taxones bacterianos transcripcionalmente activos y dominantes como *Fischerella*, fueron predados por las poblaciones virales líticas más abundantes. Mientras que, las bacterias

heterotroficas de los filos Proteobacterias y Firmicutes se asociaron a virus lisogénicos inducidos espontáneamente o por mitomicina C, respectivamente, revelando un nexo entre las funciones microbianas (metabolismo) y el tipo de ciclo infecciososo.

Finalmente, analizamos los tres metagenomas virales de Porcelana y Cahuelmó junto con los nueve metagenomas virales de aguas termales ya existentes publicados hasta la fecha, para estudiar cómo la estructura de la comunidad viral se ve afectada por los factores ambientales. El extenso análisis de las secuencias de proteínas, genes y genomas utilizando frecuencias k-mer, grupos de proteínas virales (vPC) y unidades taxonómicas operacionales virales (vOTU) mostró un patrón biogeográfico determinado por los principales factores ecológicos (pH y temperatura).

La red de proteínas compartidas de las comunidades virales termofílicas globales mostró una modularidad inesperada, lo que sugiere una restricción al flujo génico entre las fuentes termales y una alta riqueza local que se asoció a hospederos específicos al cruzar la información de espaciadores CRISPR y los módulos virales de la red. Estos análisis notaron la existencia de pares virus-hospedero específicos que permiten el mantenimiento de esta riqueza local.

En esta tesis, propusimos que las comunidades virales de los tapetes microbianos fototróficos están dominados por el orden Caudovirales, siendo los cianófagos uno de los grupos principales. Del mismo modo, las comunidades virales tienen un impacto en sus hospederos mediante interacciones líticas que estimulan la coevolución de los pares de virus-hospedero más abundantes y activos en estos sistemas térmicos, así como las interacciones lisogénicas que influyen en la adaptabilidad de los hospederos heterótrofos, probablemente a través de

conversión lisogénica. Por último, proponemos que los virus de sistemas termales terrestres siguen patrones biogeográficos donde las comunidades virales se transportan de forma pasiva por aire a escala local y global, pero luego se estructuran localmente influenciadas por las condiciones ambientales (pH y temperatura) que afectan principalmente la estructura de la comunidad de hospederos.

ABSTRACT

Viruses have proven to be ubiquitous in all environments and hot springs are not the exception even though its extreme conditions. In terrestrial hot springs with circumneutral pH and thermophilic temperatures, phototrophic mats have demonstrated for decades to be useful as models for understanding the composition, structure, and function of microbial communities in nature. Therefore, they provide a theoretical framework in which to study the viral component of these communities and the virus-host interactions.

Northern Patagonia is one of the southernmost regions of the Andean volcanic arc, and harbor 13 of the most active volcanoes in Chile. Consequently, a large number of hot springs are scattered throughout the Patagonian fjords and forests. Some examples are Porcelana and Cahuelmó hot springs located in the Comau and Cahuelmó fjords, respectively. Both hot springs are covered by microbial mats that grow along a thermal gradient between 70-46°C, dominated by bacterial oxygenic and anoxygenic phototrophs. The first group is represented by the genera *Fischerella*, *Calothrix*, *Leptolyngbya*, and *Oscillatoria*, while the anoxygenic group is represented by *Chloroflexus* and *Roseiflexus* genera. Other important taxa in these microbial mats are heterotrophic bacteria of Proteobacteria, Bacteroidetes, and Deinococcus-Termus phyla.

In the present thesis, the mining of viral sequences from cellular metagenomes and metatranscriptomes of Porcelana, shown that the viral community was predominantly

composed of Caudovirales (70%), with most of the transcriptionally active infections driven by cyanophages (up to 90% of Caudovirales transcripts). Metagenomic assembly leads to the recovery and description of the first T7-like thermophilic cyanophage (CHP58). This virus (TC-CHP58) was associated with *Fischerella* spp. by CRISPR spacers. Additionally, we have found marked differences in the number of metagenomic CRISPR loci at different temperatures and spacers diversity (different sequences and in different abundances) in *Fischerella* contigs, as well as Single Nucleotide Variants (SNVs), in the TC-CHP58 proto-spacers, which reinforce the theory of coevolution between natural virus populations and cyanobacterial hosts.

Later, using a multi-omic approach along with mitomycin C *in-situ* inductions, we studied the lytic and lysogenic viral communities of the phototrophic mats communities in Porcelana.

A high quality group of Metagenome Assembled Genomes (MAGs) obtained from Porcelana microbial mats, were interrogated for the presence of integrated temperate viruses (prophages). Also, metagenomes of natural and mitomycin C induced viral communities were analyzed to study the differential abundance of viral genomes and proteins (functions) between both viral communities, as well as ecological parameters such as α and β diversities. Our results suggest that lysogenic and lytic viral populations were strongly associated to specific hosts. The most transcriptionally active and dominant bacterial taxa such as *Fischerella*, were predated by the most abundant lytic viral populations. Meanwhile heterotrophic Proteobacteria and Firmicutes phyla were associated to spontaneously and mitomycin C induced lysogenic viruses respectively, revealing a nexus between the microbial roles (metabolism) and the type of viral infectious cycle.

Finally, we analyzed the three viral metagenomes of Porcelana and Cahuelmó together with the nine already existing hot springs viral metagenomes published to date, in order to study how viral community structure is affected by ecological drivers over the American continent in a latitudinal scale. The extensive analysis of protein, gene and genome sequences using *k-mer* frequencies, viral Protein Clusters (vPCs) and viral Operational Taxonomic Units (vOTUs) showed a biogeographic pattern according to major ecological drivers (pH and temperature).

Protein sharing network of the global thermophilic viral communities, showed an unexpected modularity, that suggests a restriction to gene flow between hot springs and a high local richness that was associated to specific hosts when crossing CRISPR spacers information and viral modules of the network. The latter notice the existence of specific virus-host pairs that allow the maintenance of this local richness.

In these thesis we propose that viral communities of phototrophic microbial mats are dominated by Caudovirales order being the cyanophages one of the major groups. Likewise, the viral communities have an impact on their hosts by lytic interactions that stimulate the coevolution of the most active and abundant host-virus pairs in these thermal systems, as well as lysogenic interactions that influence the fitness of the heterotrophic hosts, probably through lysogenic conversion. Lastly, we propose that hot spring viruses followed biogeographic patterns where viral communities are transported passively by air on a local and global scale, but then locally structured influenced by the environmental conditions (pH and temperature) that primarily affect the structure of the host community.

GENERAL INTRODUCTION

1. Terrestrial hot spring.

The terrestrial hot springs correspond to zones where subterranean aquifers heated by the action of geothermal energy emerge to the surface. They present a global distribution and are associated with areas of convergence and subduction of tectonic plates (Spiess et al., 1980). United States is the most active geothermal country in the world followed by Russia and Chile. Yellowstone National Park (YNP), situated in Wyoming (USA), is one of the best studied sites due to their extension and different thermal features having one of the highest concentrations of hot springs in the world (Zablocki et al., 2018).

Temperatures in hot springs range from 40 °C to 98 °C, and according to this are classified as moderately thermophilic (40–71 °C) or hyperthermophilic (72–98 °C) (Bell, 2012). Hot spring pH usually ranges from 1 to 10, defining three categories acidic (pH 1–5), circumneutral (pH 6–7.5) and alkaline (pH >7.5) (Zablocki et al., 2018). These features (temperature and pH) are some of the most relevant abiotic determinants in the microbial community structure in these extreme ecosystems (Brock, 1967).

2. Microbial communities of terrestrial hot springs.

Hot springs represent discontinuous habitats with a rich interphase between aquatic and terrestrial environments, which determines their physicochemical properties. Usually they

present geochemical and physical gradients distributed along contrasting distances, with scales that can go from the centimeter to hundreds of meters and kilometers (Papke et al., 2003; Sharp et al., 2014). Microbial communities from these high temperature habitats are usually dominated by few types of microorganisms (some phyla) and usually are less diverse than lower temperature freshwater habitats and oceans (Inskeep et al., 2010, 2013).

The relative simplicity of hot spring communities, has allowed its use as models to correlate genomic functions with environmental parameters, and for understanding environmental determinants over community structure (Coman et al., 2013; Inskeep et al., 2013; Klatt et al., 2013; Sharp et al., 2014). In the same way, because hot springs are a discontinuous environment that can be considered as hot islands in a cold ocean world, which acts as an effective environmental filter, they have been used to test the Louren Baas Becking's hypothesis that "everything is everywhere, but the environment select" (O'Malley, 2008).

Major drivers of bacterial and archaeal communities structure in these environments are temperature (Sharp et al., 2014), pH (Inskeep et al., 2010, 2013; Power et al., 2018) and sulfide, or elemental sulfur (Inskeep et al., 2010, 2013; Menzel et al., 2015). In general, sites with pH values between 5-9 and a temperature range of 45-70 ºC are dominated by phototrophic organisms either oxygenic or anoxygenic, depending on sulfide or elemental sulfur concentrations levels (Alcamán-Arias et al., 2018; Inskeep et al., 2013; Menzel et al., 2015). In the same range of pH, but temperatures > 70 ºC and associated to high concentrations of dissolved sulfide or elemental sulfur, the dominant groups are Aquificales and Thermoproteales (Inskeep et al., 2013; Menzel et al., 2015). Finally, Sulfolabales

members dominate the community at acidic pH range (2-5) and temperatures > 70 ºC (Klatt et al., 2013; Menzel et al., 2015).

3. Phototrophic microbial mat in terrestrial hot spring.

Microorganisms of terrestrial hot springs with circumneutral pH and moderately thermophilic temperatures usually form dense communities of stratified distribution called microbial mats, located at the interface formed between the substrates of soil and water (Miller et al., 2009). Frequently, the uppermost layer of the mat is composed of photoautotrophs; such as oxygenic phototrophic cyanobacteria, including the unicellular cyanobacterium *Synechococcus* spp. (Bhaya et al., 2007; Klatt et al., 2013; Thiel et al., 2016), the filamentous non-heterocystous *Oscillatoria* spp., the filamentous heterocystous *Fischerella* spp. (Alcamán-Arias et al., 2018; Alcamán et al., 2015; Alcorta et al., 2018; Mackenzie et al., 2013; Miller et al., 2006), as well as filamentous anoxygenic phototrophs (FAPs), such as *Roseiflexus* sp. and *Chloroflexus* sp. (Klatt et al., 2013; Thiel et al., 2016). These primary producers interact with heterotrophic prokaryotes through element and energy cycling (Alcamán-Arias et al., 2018; Klatt et al., 2013; Thiel et al., 2016).

Furthermore, other phyla such as Deinococcus-Thermus, Proteobacteria, Firmicutes, and Bacteroidetes are also prominent in these mats (Alcamán-Arias et al., 2018; Klatt et al., 2013; Thiel et al., 2016). Members of Deinococcus-Thermus such as Meiothermus are known chemoheterotrophic aerobic bacteria with the capacity to degrade complex carbon sources (Klatt et al., 2013; Thiel et al., 2016). Some members of class Deltaproteobacteria are known to have a sulfate-reducing metabolism in these mats, while representatives of the

Alphaproteobacteria are known to carry BChla and then the potential of perform chlorophototrophy (Thiel et al., 2016). Finally, Firmicutes and Bacteoridetes role in these communities may involve fermentation and the degradation of complex carbon compounds (Klatt et al., 2013).

These commonly simplified but highly cooperative communities have been historically used as models for understanding the composition, structure, and function of microbial consortia (Klatt et al., 2011, 2013). Generally, in microbial mats, the organic matter produced by phototrophs is subsequently used as the main source of energy and organic carbon by aerobic and anaerobic heterotrophic microorganisms. Later, the aerobic heterotrophs play an important role for the community because their activity leads to oxygen depletion, providing an anoxygenic environment. This condition is required by fermentative organisms which subsequently provide growth substrates to the sulfate-reducing bacteria. Other minority groups are composed by nitrifying and denitrifying bacteria, and also methanogenic bacteria (Gemerden, 1993). Cyanobacteria as primary producers are particularly relevant in many microbial mats (Alcamán-Arias et al., 2018; Alcamán et al., 2015; Bhaya et al., 2007; Miller et al., 2006) because they can combine CO_2 and N_2 fixation, the two most important biogeochemical processes on the Earth (Klatt et al., 2011).

Additionally the role of abiotic factors, such as pH, sulfide concentration, and temperature, in determining microbial assemblages and life cycles of these communities have been extensively investigated in phototrophic mats (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Cole et al., 2013; Inskeep et al., 2013).

4. Viruses in natural microbial communities.

In nature, prokaryotic organisms must deal with a strong predation pressure, mostly of viral origin (Rodriguez-Valera et al., 2009), therefore viruses are essential components of microbial communities contributing to their fitness and evolution, trough their infective cycle (Howard-Varona et al., 2017; Suttle, 2007; Wilhelm and Suttle, 2000). In these terms, cellular microbial diversity is affected not only by the availability of resources and physical conditions but also by the viral fraction of the microbial community. In this way, phages (viruses that infect bacteria) influence global biogeochemical cycles by cell lysis. It has been estimated that viruses can kill between 4–50% of the bacteria produced every day in the oceans (Breitbart and Rohwer, 2005). In environments such as marine ecosystems, the effects of viruses on nutrient cycles have been quite studied (Hewson et al., 2010; Suttle, 2007; Wilhelm and Suttle, 2000). Therefore, is known that lysis involves the release of organic carbon and other nutrients to the environment where they are recycled, in a process known as "Viral Shunt" (Rohwer and Thurber, 2009; Wilhelm and Suttle, 2000). Organic carbon in marine systems can be separated into dissolved organic carbon (DOC) and particulate organic carbon (POC) (Wilhelm and Suttle, 1999). These two carbon stocks behave differently, the majority of DOC is recycled in the microbial loop by heterotrophic bacteria and it is not transferred to higher trophic levels, while the majority of the POC is transferred directly to higher trophic levels by its consumption by micro herbivores (Wilhelm and Suttle, 1999). This is how viral lysis diverts the carbon from the POC to the DOC reservoir producing a "short circuit" in the microbial loop and controlling the type of organic carbon available in the water column (Danovaro et al., 2008; Wilhelm and Suttle, 1999).

It has been described that viral lysis and subsequent generation of DOC can increase bacterial abundance up to 10 times (Gobler et al., 1997), which leads to an increase in bacterial production, but a decrease in carbon transfer to higher trophic levels (Hewson et al., 2010).

The lifestyles of viruses have been described to follow three specific paths, lytic (productive), lysogenic and pseudolysogenic (Miller and Day, 2008). Lytic or productive infection is characterized by a rapid replication of phage genome, transcription and translation of the viral components to finally release viral particles by cell lysis (Miller and Day, 2008). Lysogeny and pseudolysogeny are characterized by the persistence of the viral genome inside the host cell but without the production of viral particles and then not produce the cell death by lysis. Both states differ in the stable integration of the viral genome into the chromosome of the host, the latter process does not occur in the pseudolysogeny, where the viral genome remains unstable, i.e., circularized as a low copy plasmid, and usually failing in its replication as productive infection or when the cell is dividing (Miller and Day, 2008).

The type of viral lifestyle have different effects in the ecology of the host. Lytic infections usually lead to coevolution in an aggressive evolutionary 'arms race' in which viruses and host constantly evolve mechanisms of resistance to each other (Rohwer and Thurber, 2009). Lytic viruses in this way, directly influence the diversity of their host by selectively lysing the dominant taxa, which usually are the most active one in the microbial communities (Suttle, 2007). Consequently, the lytic lifestyle has been correlated to the ecological model of "Kill the Winner" (Knowles et al., 2016, 2017; Rohwer and Thurber, 2009). Kill the Winner (KtW) hypotheses of lytic infection predict that viral predation efficiency is dependent of the activity and density of the host, and consequently to the frequency of host-virus encounters. In this

sense, blooms of rapidly growing hosts (winners) will be terminated (kill) by the action of phages, increasing host diversity by the arise of natural resistance (Knowles et al., 2016; Thingstad et al., 2008).

Lysogenic and pseudolysogenic infections in turn, leads to a symbiotic relation with their host, forming a new biological entity known as lysogen. Lysogenic cycle brings benefits to hosts, that includes prophage mediated immunity against other virus infections, protection from grazers predation by the acquisition of new virulence factors and gain of new metabolic functions through transduction (Feiner et al., 2015; Howard-Varona et al., 2017; Knowles et al., 2017). Temperate viruses can also influence the microbial communities when they enter in a productive cycle by lysing competitor strains or lysogenizing other microorganisms (Howard-Varona et al., 2017). Also they can produce cooperative effects in the communities by liberating intracellular contents for neighboring cells to be used as nutrients (Howard-Varona et al., 2017). Therefore, lysogenic cycle fits better to other ecological model where viruses will seek to establish a symbiotic relationship with their host lysogenizing the most active and usually dominant taxa in the microbial community, following the proposed "Piggyback the Winner" (PtW) model (Knowles et al., 2016, 2017).

Lytic and lysogenic dynamics have been studied mostly in aquatic environments (marine and freshwater) with a few studies focused in sediments and soils (Howard-Varona et al., 2017; Knowles et al., 2017), reaching to the overall conclusion that lysogeny is a common viral strategy in all the ecosystems (Howard-Varona et al., 2017; Knowles et al., 2017).

Viruses have been observed to be globally distributed and have high diversity on the local scale (Breitbart and Rohwer, 2005; Brum et al., 2015). These observations can be explained by

the seed bank model, where only the most abundant viruses are active, whereas the rest of the viruses are inactive and then rare, forming a potential population (bank) that is resisting until the right moment arrives (Breitbart and Rohwer, 2005). This model matches the abundance curves observed in marine environments, where the most abundant viral genome usually does not exceed the 5% of the total community and most of the viral genomes are extremely rare, usually ≤ 0.01% of the community (Breitbart and Rohwer, 2005; Brum et al., 2015). When environmental conditions change, a different hosts grow faster and the viruses infecting these hosts move from the reservoir "bank" into the "active" fraction, and the other way round (Breitbart and Rohwer, 2005).

5. Viral communities in terrestrial hot springs.

Viruses have proven to be ubiquitous, numerous and active components of the hot springs microbial communities (Bolduc et al., 2012, 2015; Menzel et al., 2015; Munson-Mcgee et al., 2018; Schoenfeld et al., 2008; Sharma et al., 2018; Zablocki et al., 2017). They constitute the major biotic factor that can shape the diversity of their host through coevolution (Sano et al., 2018), and regulating the structure of cellular communities through predation in hot springs (Breitbart et al., 2004; Klatt et al., 2013; Pride and Schoenfeld, 2008; Schoenfeld et al., 2008). Most viral investigations carried out in hot springs, occur within the source waters and usually considering only the lytic populations (Bolduc et al., 2012; Rachel et al., 2002; Schoenfeld et al., 2008; Wang et al., 2015; Zablocki et al., 2017). In these thermal waters, virus abundances range between 10^4 and 10^9 Virus Like Particles (VLP) mL^{-1}, with a virus to bacteria ratio (VBR) of 5 (Breitbart et al., 2004; Redder et al., 2009; Schoenfeld et al., 2008). They play an

important role in the structure of host populations and as drivers of organic and inorganic nutrient recycling (Breitbart et al., 2004). The majority of the thermophilic viruses described to date are dsDNA, with new and complex viral morphotypes, distinct to the typical head and tail morphologies (Pawlowski et al., 2014; Prangishvili and Garrett, 2004; Rachel et al., 2002; Redder et al., 2009; Schoenfeld et al., 2008). Furthermore, the few metaviromes obtained in thermal waters indicate that natural thermophilic viral communities differ from those obtained in culture, given that there was only a 20-50% similarity between the sequences obtained compare to those in the databases (Bolduc et al., 2015; Diemer and Stedman, 2012; Pride and Schoenfeld, 2008; Schoenfeld et al., 2008; Zablocki et al., 2018)

Database dependent analyses from these studies, reported that in general the structure of viral communities at hyperthermophilic, circumneutral pH hot springs were dominated by bacterial and Termoproteales viruses from Myoviridae, Siphoviridae, Podoviridae and Globuloviridae families (Pride and Schoenfeld, 2008; Schoenfeld et al., 2008). Meanwhile at acidic pH and hyperthermophilic temperatures, Sulfolobales viruses were the dominant and usually associated to a single viral family such as Lipothrixviridae, Fuselloviridae, Ampullaviridae, Bicaudaviridae and Rudiviridae (Bolduc et al., 2015; Gudbergsdóttir et al., 2016; Menzel et al., 2015). For instance, viral communities from thermophilic sites (40 - 71ºC) and circumneutral pH have reported that Caudovirales order, dominate in these type of hot springs showing that the most abundant viruses would infect Cyanobacteria and the Planctomycetes genus *Gemmata* (Zablocki et al., 2017). Genomic analyses of the recovered cyanophage sequences showed that those viruses were close to freshwater cyanophages that infects filamentous cyanobacteria (Zablocki et al., 2017).

Even when, lysogeny have been considered an effective lifestyle in hot springs (Breitbart et al., 2004; Schoenfeld et al., 2008; Sharma et al., 2018), this strategy have not been systematically studied in these environments, and most of the studies have only randomly found prophages (Sharma et al., 2018) or molecular markers related to that lifestyle such as integrases (Schoenfeld et al., 2008).

7. Northern Patagonia terrestrial hot springs.

The Patagonian Andes (39 – 47° South latitude) are a volcanically active region and therefore, many hot springs are scattered throughout the fjords and forests. The region is covered by Valdivian temperate rainforest, a unique ecosystem considered to be a hotspot for biodiversity of plants and vertebrates (Arroyo *et al*., 2004). Porcelana hot spring is located at the Comau fjord in the Huequi peninsula, in an area of shallow depth geothermic events whose outflow pours over volcanic rocks originated in the Quaternary Period, and that are encircled by active Quaternary volcanoes such as the Huequi Volcano (Duhart *et al*., 2000). The volcanic rocks in this region are formed mainly by silicates and carbonates, and are rich in metallic minerals and elements such as pyrite, chalcopyrite, arsenopyrite, and Antimony (Fortey *et al.,* 1992).

Cahuelmó hot spring is located across the Comau fjord, at 23.1 km from Porcelana hot spring in the coast of Cahuelmó Fjord (at sea level), exposed to brackish water influence (salinity 3%) and winds. This system is located in a metamorphic rock complex, closer to the Andes mountain chain, so it is rich in metallic minerals and elements such as pyrite, polonium, magnetite, and chalcopyrite (Duhart *et al*.,2000).

Porcelana and Cahuelmó hot springs are covered by microbial mats that grow along a thermal gradient between 70-46°C, dominated by bacterial phototrophs, such as filamentous oxygenic phototrophs of the genus *Fischerella, Calothrix*, *Leptolyngbya* and *Oscillatoria* plus filamentous anoxygenic phototrophs of *Chloroflexus* and *Roseiflexus* genera (Alcamán-Arias et al., 2018; Mackenzie et al., 2013). Other important taxa in these microbial mats are heterotrophic bacteria of Proteobacteria, Bacteroidetes, and Deinococcus-Termus phyla (Alcamán-Arias et al., 2018; Mackenzie et al., 2013). Proteobacteria are mainly represented by genera *Halomonas* and *Shewanella* of class Gammaproteobacteria, while Bacteroidetes are represented mainly by members of the Sphingobacteriales order (Mackenzie et al., 2013). Finally Deinoccus-Thermus are represented mostly by *Meiothermus* genus (Alcamán-Arias et al., 2018; Mackenzie et al., 2013).

8. Viruses in Northern Patagonia hot springs. Why study them there?

Environmental viruses entered to the game of microbial ecology late, however 40 years have passed since their presence in marine waters was noticed for the first time on the coasts of Oregon, USA (Torrella and Morita, 1979), and a decade later quantitative estimates revealed that each milliliter of seawater contains millions of these particles (Bergh et al., 1989).

Again, more than a decade passed until a fraction of the billions of particles present in the ocean were sequenced for the first time by Mya Breitbart a former Ph.D. student of Forest Rohwer in San Diego State University (Breitbart et al., 2002).

Chile, entered into the arena of environmental virology in 2011 when Jean-Michel Claverie isolated the largest giant virus existing to date, the *Megavirus chilensis* in the protected area of

the Coastal Marine Research Station (ECIM-PUC) (Arslan et al., 2011). Eight years have passed since this milestone for environmental virology in Chile. However, the study of environmental viruses in our country has been evasive and it has not attracted many followers, mainly due to ignorance of its existence and the few national research groups in the area.

The geographical location of Chile not only generates a country full of contrasts with a coast of ~ 5000 km, but also positions us as a polyextremophile country. Chile harbors the driest hot desert in the world (Atacama desert), the third largest geothermal field in the world and first in South America (El Tatio), but also the "Andean volcanic arc" which includes over 200 potentially active volcanoes (Stern, 2004). Northern Patagonia is the southern most region of the Southern Volcanic Zone of the arc (SVZ) and includes 13 of the more active volcanoes reported in Chile (Stern, 2004).

The active volcanism of the fjords region in Patagonia and the geographical isolation of its landscapes, has allowed us to propose the present doctoral thesis where our main aim is to understand in its broadest sense the viral communities that prey in the natural communities of phototrophic mats in the pristine hot springs of northern Patagonia. We have chosen this study model because thermophilic phototrophic mats have demonstrated for decades to be useful for understanding the composition, structure, and function of microbial communities in nature. Therefore, they provide a theoretical framework in which to study the viral component of these communities. Although aspects of the ecological impacts and abundance of viruses in themophilic phototrophic mats has been investigated (Davison et al., 2016; Heidelberg et al., 2009), currently, the diversity of these viruses remains excessively undersampled. Therefore, this lack of knowledge gives us the opportunity to use the hot springs of Patagonia or many

others scattered throughout our country as a model system in which to study these unknown viral communities.

HYPOTHESIS

Viral communities of phototrophic mats are genetically diverse, and have an impact in the diversity and evolution of their cellular host in terrestrial hot springs.

GENERAL AIM

Characterize the structure, activity, lifestyle, and putative host of the thermophilic viral communities in phototrophic microbial mats and compare them at local and global scale.

SPECIFIC AIMS

1. To determine the structure, activity, and putative hosts of the thermophilic viral community through metagenomic and metatranscriptomic mining of viral sequences in Porcelana phototrophic microbial mats.

2. To identify and compare lytic, and lysogenic viral communities from natural and mitomycin C induced phototrophic microbial mats, through viral metagenomics in Porcelana hot spring.

3.To determine and compare the genomic and structural variability of viral communities on a local (Patagonia) and global scale by viral metagenomics and database independent methods.

CHAPTER 1

Active crossfire between Cyanobacteria and Cyanophages in phototrophic mat communities within hot springs.

Active crossfire between Cyanobacteria and Cyanophages in phototrophic mat communities within hot springs

Sergio Guajardo-Leiva[1], Carlos Pedrós-Alió[2], Oscar Salgado[1], Fabián Pinto[1], and Beatriz Díez[1,3]

[1]Department of Molecular Genetics and Microbiology, Pontificia Universidad Católica de Chile, Santiago, Chile.

[2]Programa de Biología de Sistemas, Centro Nacional de Biotecnología (CSIC), Madrid, España.

3Center for Climate and Resilience Research (CR)2, Chile.

*** Correspondence:**
Beatriz Díez
bdiez@bio.puc.cl

Hot-springs, Cyanophages, Phototrophic Microbial Mat, CRISPR , Thermophilic Cyanobacteria.

ABSTRACT

Cyanophages are viruses with a wide distribution in aquatic ecosystems, that specifically infect Cyanobacteria. These viruses can be readily isolated from marine and fresh waters environments; however, their presence in cosmopolitan thermophilic phototrophic mats remains largely unknown. This study investigates the morphological diversity (TEM), taxonomic composition (metagenomics), and active infectivity (metatranscriptomics) of viral communities over a thermal gradient in hot spring phototrophic mats from Northern Patagonia (Chile). The mats were dominated (up to 53%) by cosmopolitan thermophilic filamentous true-branching cyanobacteria from the genus *Mastigocladus*, the associated viral community was predominantly composed of Caudovirales (70%), with most of the active infections driven by cyanophages (up to 90% of Caudovirales transcripts). Metagenomic assembly lead to the first full genome description of a T7-like Thermophilic Cyanophage recovered from a hot spring (Porcelana Hot Spring, Chile), with a temperature of 58ºC (TC-CHP58). This could potentially represent a world-wide thermophilic lineage of podoviruses that infect cyanobacteria. In the hot spring, TC-CHP58 was active over a temperature gradient from 48 to 66ºC, showing a high population variability represented by 1979 Single Nucleotide Variants (SNVs). TC-CHP58 was

associated to the *Mastigocladus* spp. by CRISPR spacers. Marked differences in metagenomic CRISPR *loci* number and spacers diversity, as well as SNVs, in the TC-CHP58 proto-spacers at different temperatures, reinforce the theory of co-evolution between natural virus populations and cyanobacterial hosts. Considering the importance of cyanobacteria in hot spring biogeochemical cycles, the description of this new cyanopodovirus lineage may have global implications for the functioning of these extreme ecosystems.

INTRODUCTION

Hot springs host microbial communities dominated by a limited variety of microorganisms that form well-defined mats (Uldahl and Peng, 2013; Inskeep *et al*., 2013). Frequently, the uppermost layer of the mat is composed of photoautotrophs; such as oxygenic phototrophic cyanobacteria, including the unicellular cyanobacterium *Synechococcus* spp. (Steunou *et al*., 2006, 2008; Bhaya *et al*., 2007; Klatt *et al*., 2011), the filamentous non-heterocystous *Oscillatoria* spp., the filamentous heterocystous *Mastigocladus* spp. (Stewart, 1970; Miller *et al*, 2006; Mackenzie *et al*., 2013, Alcamán *et al*., 2015), as well as filamentous anoxygenic phototrophs (FAPs), such as *Roseiflexus* sp. and *Chloroflexus* sp. (van der Meer *et al*., 2010; Liu *et al.,* 2011; Klatt *et al*., 2011). These primary producers interact with heterotrophic prokaryotes through element and energy cycling (Klatt *et al*., 2013). Heterocystous cyanobacteria are a key component in hot springs, since these systems are commonly N-limited due to the rapid assimilation and turnover of inorganic nitrogen forms (Lin *et al*., 2015; Alcamán *et al*., 2015). Thus, N_2-fixation by cyanobacteria is identified to be a key biological process in neutral hot spring microbial mats (Alcamán *et al*., 2015).

These simplified but highly cooperative communities have been historically used as models for understanding the composition, structure, and function of microbial consortia (Klatt *et al*., 2011; Inskeep *et al*., 2013). The role of a variety of abiotic factors, such as pH, sulfide concentration, and temperature, in determining microbial assemblages and life cycles in these ecosystems have been investigated (Inskeep *et al*., 2013; Cole *et al*., 2013). However, there is a lack of investigation into biotic factors, such as viruses, on thermophilic photoautotrophic mats, with existing studies only reporting short or partial viral sequences, (Heidelberg *et al*., 2009; Davison *et al*., 2016). Currently, viral communities from thermal mats have been characterized through indirect approaches, indicating

the hypothetical presence of viruses (Heidelberg *et al*., 2009; Davison *et al*., 2016). Heidelberg *et al.* (2009) used CRISPR spacer sequences extracted from the genomes of two thermophilic *Synechococcus* isolates, from a phototrophic mat in Octopus Spring. Subsequently, they searched for viral contigs from previously published water metaviromes from the Octopus and Bear Paw Springs in Yellowstone National Park (USA) (Schoenfeld *et al.*, 2008). Furthermore, Davison *et al.* used CRISPR spacers and nucleotide motive frequencies to link viral contigs to known hosts using a metavirome obtained by Multiple Displacement Amplification (MDA) of VLPs from a mat in Octopus Spring (Davison *et al*., 2016), as well as reference genomes from dominant species (*Synechococcus* sp., *Roseiflexus* sp. and *Chloroflexus* sp.) previously described in the same microbial mat. A key finding from these studies was the link between viruses and their hosts, indicating their co-evolution and an effective "arms race" within hot spring phototrophic mats.

Unlike thermophilic mat studies, most viral investigation carried out in hot springs occur within the source waters (Rachel *et al*., 2002; Yu *et al*., 2006; Schoenfeld *et al.*, 2008; Bolduc e*t al*., 2012; Bolduc *et al*., 2015; Zablocki *et al*., 2017). In these waters, virus abundances range between 10^4 and 10^9 Virus Like Particles (VLP) mL^{-1} (Breitbart *et al*., 2004; Schoenfeld *et al*., 2008; Redder *et al*., 2009). They play an important role in both the structuring of host populations and as drivers of organic and inorganic nutrient recycling (Breitbart *et al*., 2004). The majority of the viruses were dsDNA, with new and complex viral morphotypes, distinct to the typical head and tail morphologies (Rachel *et al*., 2002; Prangishvili and Garrett, 2004; Schoenfeld *et al*., 2008; Redder *et al*., 2009; Pawlowski *et al*., 2014). Furthermore, the few metaviromes obtained in thermal waters indicate that natural thermophilic virus communities differ from those obtained in culture, given that there was only a 20-50% similarity between the sequences obtained and those in the databases (Pride and Schoenfeld, 2008; Schoenfeld *et al*., 2008; Diemer and Stedman, 2012; Bolduc *et al*., 2015). Thus far, the genomes that have been isolated and sequenced from thermophilic viruses (57 genomes, of which 37 infected archaea and 20 infected Bacteria) generally yielded few significant matches to sequences in public databases (Uldahl and Peng, 2013). More recently, a water metaviromic study from Brandvlei hot spring (BHS), South Africa (Zablocki *et al*., 2017) reported the presence of two partial genomes (10 kb and 27 kb), the first related to Podoviridae and the second to lambda-like Siphoviridae families. Both Caudovirales genomes did not have a confirmed host, but the presence of green microbial mat-patches around the contours of the hot spring, implied that filamentous Cyanobacteria and unclassified *Gemmata* species

were the potential hosts, respectively. The last, based on the proximity of some viral predicted proteins with bacteria from well characterized microbial mats present in a nearby hot spring (Tekere *et al*., 2001; Jonker *et al*., 2013).

Given the lack of knowledge of viral communities within hot spring phototrophic microbial mats, the present study used the mats of Porcelana hot spring (Northern Patagonia, Chile), as a pH neutral model, to better understand the associated thermophilic viral communities within these mats. This pristine spring is covered by microbial mats that grow along a thermal gradient between 70-46°C, dominated by bacterial phototrophs, such as filamentous cyanobacteria from the genus *Mastigocladus* (Mackenzie *et al*., 2013; Alcamán *et al*., 2015). This is the dominant and most active cyanobacterial genus in the Porcelana mat environment, carrying out important biological processes such as carbon- and N_2-fixation (Alcamán et al., 2015, 2017). Thus, this study proposes that the mats in Porcelana hot spring are dominated by viral communities of the Order Caudovirales, which is able to infect Cyanobacteria, preferably *Mastigocladus* spp.

The viral diversity in Porcelana was determined through the detection of viral signals in microbial mat omics data, and by TEM along the thermal gradient. The results demonstrate that the viral community was dominated by Caudovirales, which actively infect Cyanobacteria. Furthermore, the first complete genome description of a thermophilic cyanobacterial T7-like podovirus, Thermophilic Cyanophage Chile Porcelana 58°C (from now on TC-CHP58) is realized. The host is the dominant phototroph *Mastigocladus* spp, based on CRISPR spacers. Finally, the presence of different populations of this new podovirus are identified through single nucleotide variants (SNVs) analyses, and the co-evolution of *Mastigocladus* spp. and particular populations of TC-CHP58 at different temperatures is described through association of specific SNVs to different CRISPR spacers.

MATERIAL AND METHODS

Sampling site.

Porcelana hot spring is located in Chilean Patagonia (42º 27’ 29.1’’S - 72º 27’ 39.3’’W). It has a neutral pH range between 7.1 to 6.8 and temperatures ranging from 70 ºC to 46 ºC, when sampled on March 2013. Phototrophic microbial mats growing at 66 ºC, 58 ºC and 48 °C were sampled using a cork borer

of 7 mm diameter. Cores of 1 cm thick were collected in triplicate at noon (12:00 PM), transported in liquid nitrogen and kept at -80 ºC until DNA and RNA extraction.

Transmission electron microscopy.

Five liters of interstitial fluid was squeezed using 150 µm sterilized polyester net SEFAR PET 1000 (Sefar, Heiden, Switzerland) and filtered through 0.8 µm pore-size polycarbonate filters (Isopore ATTP, 47 mm diameter, Millipore, Millford, MA, USA) and 0.2 µm pore-size (Isopore GTTP, 47 mm diameter, Millipore) using a Swinex filter holder (Millipore). Particles in the 0.2 µm filtrate were concentrated to a final volume of approximately 35 mL using a tangential-flow filtration cartridge (Vivaflow 200, 30 kDa pore size, Vivascience, Lincoln, UK). Viral concentrates (15 µL) were spotted onto Carbon Type-B, 200 mesh, Copper microscopy grids (Ted Pella, Redding, California, USA), stained with 1% uranyl acetate and imaged on an FEI Tecnai T12 electron microscope at 80 kV (FEI Corporate, Hillsboro, Oregon, USA) with attached Megaview G2 CCD camera (Olympus SIS, Münster, Germany). Imaging analysis was done at the Advanced Microscopy Unit, School of Biological Sciences at Pontificia Universidad Católica de Chile (Santiago, Chile).

Nucleic acid extractions and high throughput sequencing.

Nucleic acids (DNA and RNA) were extracted as previously described (Alcamán et al., 2015). For RNA, Trizol (Invitrogen, Carlsbad, California, USA) was added to the mat sample, and homogenized by bead beating, two pulses of 20 seconds. Quality and quantity of the extracted nucleic acids were checked and kept at -80 °C.

Samples were sequenced by Illumina Hi-seq technology (Research and Testing Laboratory, Texas, USA). Briefly, for metagenomes, DNA was fragmented using NEBNext dsFragmentase (New England Biolab, Ipswich, Massachusetts, USA), followed by DNA clean up using column purification, and a NEBUltra DNA Library Prep Kit for Illumina (New England Biolab, Ipswich, Massachusetts, USA) was used for library construction.

For metatranscriptomes, DNase treated total RNA was cleaned up of rRNA by a Ribo-Zero rRNA Removal Kit Bacteria (Illumina, San Diego. California, USA), followed by purification using an Agencourt RNAClean XP Kit (Beckman Coulter, Indianapolis, Indiana, USA), and a NEXTflexTM Illumina Small RNA Sequencing Kit v3 (Bio Scientific, Austin, Texas, USA) was used for library construction.

For quality filtering, the following filters were applied using Cutadapt (Martin, 2011), leaving only mappable sequences longer than 30 bp (-m 30), with a 3' end trimming for bases with a quality below 28 (-q 28), a hard clipping of the first 5 leftmost bases (-u 5), and finally a perfect match of at least 10 bp (-O 10) against the standard Illumina adaptor. Finally, the removal of sequences representing simple repetitions that are usually due to sequencing errors was applied using PRINSEQ (Schmieder and Edwards, 2011) DUST threshold 7 (-lc_method dust, -lc_threshold 7). Details of the number of sequences obtained are shown in Supplementary Table S1.

Identification of rRNA-like sequences and viral mining from metagenomes and metatranscriptomes.

Metagenomic Illumina TAGs (miTAGs) (Logares et al., 2014) that are small subunit (SSU) 16S and 18S rRNA gene sequences in the metagenomes were identified and annotated using the Ribopicker tool (Schmieder et al., 2012) with the Silva 123 SSU database (Quast et al., 2013).

For viral mining, bacterial, archaeal and eukaryotic sequences were removed through end-to-end mapping, allowing a 5% of mismatch (-N 1 -L 20) against the NCBI non-redundant (NR) database (Nov-2015) using bowtie2 (Langmead and Salzberg, 2012). Viral sequences were then recruited against modified NCBI RefSeq (Release 75) viral proteins, where only amino acid sequences from viruses that do not infect animals (NAV) were considered to build the database, using the UBLAST algorithm (-strand both -accel 0.9) through the USEARCH sequence analysis tool (Edgar, 2010). Recruitment was made for sequences with over 65% of coverage and an E-value < 1 x 10-3 (-query_cov 0.65 -evalue 1e-3). For taxonomic assignment, recruited sequences were aligned against the NAV database using BLASTX (Camacho et al., 2009) and parsed using the lowest common ancestor algorithm trough MEGAN 6 (Huson et al., 2016) (LCA score =30). The latter displays a graphical representation of

abundance for each taxonomic group identified at the family and species levels. Species classification of viral reads, was used to infer the phyla of the putative hosts based on viral RefSeq host information or through a manual search of the publication associated with each viral genome.

To extract putative viral genomes, all metagenomes (48 °C, 58 °C and 66 °C) were assembled using De Bruijn graphs as implemented in the Spades assembler (Bankevich et al., 2012), followed by gene prediction using Prodigal software (Hyatt et al., 2010) and the recovery of circular contigs over 5 kb using a Python script (Crits-Christoph et al., 2016). Only sequences over 5 kb were used in the subsequent analysis because all dsDNA viruses in the databases have genomes over that size. A homology search of the viral predicted proteins by Prokka (Seemann, 2014) was done using BLASTX against the NAV protein database and NCBI nr as described before. Additionally, all contigs over 5 kb were analyzed using VirSorter (Roux et al., 2015a) against the virome database option.
To quantify the abundance and activity of the retrieved viral genome, reads recruitment from each metagenome and metatranscriptome was performed using BWA-MEM (-M), resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/).

Phylogenetic analysis.

The protein inferred sequences of DNA polymerase and major capsid were aligned by Muscle (Edgar, 2004) and MAFFT (Katoh et al., 2002) respectively, using the amino acid substitution model determined by ProtTest 3 (Blosum62+G+F) (Darriba et al., 2011) and modelFinder (LG+F+G4) respectively. The Bayesian Markov chain Monte Carlo method was implemented with MrBayes 3.6 (Ronquist et al., 2012) and MCMC results were summarized with Tracer 1.6 (http://beast.bio.ed.ac.uk/Tracer). MrBayes was run using two independent runs, four chains, 1,500,000 generations and a sampling frequency of 100 with a burn-in value of 33% until the standard deviations of split frequencies remained below 0.01.

The maximum likelihood method was implemented with IQtree (-bb 10000 -nm 10000 -bcor 1 -numstop 1000) (Trifinopoulos et al., 2016) using 100 standard bootstrap and 10,000 ultrafast bootstrap to evaluate branch supports. The details of the sequences used for phylogenetic analyses are listed in Supplementary Table S2.

CRISPR/Cas virotopes.

Assemblies for each temperature, were taxonomically grouped (bins) using the Expectation–Maximization (EM) algorithm implemented in MaxBin 2.0 (Wu et al., 2016). In order to asses the completeness and contamination of each bin, CheckM (Parks et al., 2015) analyses were performed. Finally, the closest genome of each bin was searched using the Tetra Correlation Search (TCS) analysis implemented in Jspecies tool (Richter et al., 2016) with selection criteria of Z score greater than 0.999 and ANI over 95% (Konstantinidis et al., 2017).

CRISPR/Cas *loci* were identified in contigs assigned to *Mastigocladus* spp. from 48, 58 and 66 ºC assembled metagenomes using CRISPRFinder tool (Grissa et al., 2007). To quantify the activity of the CRISPR *loci*, reads recruitment from metatranscriptomes for the same temperatures was performed using BWA-MEM (-M), and the resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/) and normalized by total number of reads and length of each *loci*.

Spacers from CRISPR containing contigs were mapped to viral contigs using bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1). Mapped spacers were manually annotated to the viral predicted proteins in viral contig.

Single Nucleotide Variants (SNV).

To call variants occurring in TC-CHP58 populations at the 3 different metagenome temperatures, LoFreq method (Wilm et al., 2012) was used. SNVs frequencies were quantified in ORFs from TC-CHP58 genome using Bedtools suite (Quinlan and Hal, 2010). The alleles of SNVs present in proto-spacers were visualized in IGV tools for each virotope at each temperature.

RESULTS

Morphological and genetic composition of VLPs.

Transmission electron microscopy (TEM) was applied to identify the VLPs present in the interstitial fluid from microbial mats in Porcelana hot spring. Caudovirus-like particles belonging to Myoviridae, Podoviridae and Siphoviridae families, typically infecting bacteria (Figure 1A to 1G) were identified. Additionally, filamentous and rod shaped VLPs were detected, that could be associated with Lipothrixviridae and Clavaviridae families, usually infecting archaea (Figure 1H, to 1K).Viral read counts ranged between 0.47% and 0.78% of the total metagenome reads, and between 0.35% and 3.71% in the metatranscriptomes (Supplementary Table S1). At all temperatures, viral metagenomic sequences (Figure 2) revealed the dominance of the Order Caudovirales, followed by the Order Megavirales, with ~70% and ~23% of the total viral reads, respectively. Metatranscriptomic analysis results (Figure 2) showed a slightly different pattern, with a reduction in Caudovirales with increasing temperature (from ~78% at 48 °C to ~57% at 66 °C), whereas Megavirales did the opposite (from ~7% at 48 °C to ~36% at 66 °C).

In the metagenomes, Siphoviridae was the most abundant family of Caudovirales, with maximum abundance at 48°C. Myoviridae members were also well represented with a maximum of ~31% at 58°C and a minimum (~25%) at 48°C. Meanwhile, Podoviridae accounted for just ~8% at all temperatures (Figure 2). In metatranscriptomes, Siphoviridae increased six-fold with temperature, while Podoviridae and Myoviridae decreased with temperature (five- and two-fold, respectively).

The Megavirales order was also present, however at a lower abundance compared to Caudovirales. Megavirales were represented by Phycodnaviridae (~13%), Mimiviridae (~8%) and Maseilleviridae (~2%) families, remaining constant through all temperatures. Metatranscriptomics showed an increase in abundance of these three virus families with temperature.

Caudovirales host assignments.

Porcelana mat communities based on miTAGs were dominated by bacteria (~96%), with low abundances of eukarya (~3%) and archaea (~1%) (Supplementary Table S1). At the phylum level

(Figure 3A), bacterial communities were mostly composed of Cyanobacteria oxygenic phototrophs (33%, 53% and 21% of total rRNA SSU sequences at 48, 58, and 66 °C, respectively) and Chloroflexi anoxygenic phototrophs (higher than Cyanobacteria only at 66°C, with 35% of total rRNA SSU sequences). Other representative members of the community were Proteobacteria (5% to 11%), Deinococcus-Thermus (2% to 7%), Firmicutes (1% to 17%) and Bacteroidetes (4% to 8%) (Figure 3A).

The host assignment, based on taxonomy from viral reads of the most representative Caudovirales (Figure 3B), showed that viruses putatively infected members of the bacterial phyla Proteobacteria, Cyanobacteria, Actinobacteria and Firmicutes. Metagenomic data showed that increases in temperature led to an increase in viruses from Actinobacteria and Firmicutes. Additionally, an increase in Cyanobacteria viruses was observed at 58°C. Viruses from Proteobacteria, Actinobacteria and Firmicutes were represented by the three Caudovirales families, while viruses from Cyanobacteria were represented by Podoviridae and Myoviridae families only (Supplementary Table S3), where cyanopodovirus and cyanomyovirus reads increase from 31% to 50% at 48ºC and from 30% to 45% at 58ºC, then decrease to 23% and 28% at 66ºC, respectively.

Metatranscriptomic sequences from Caudovirales potentially infecting Cyanobacteria, were predominant at 48 °C and 58 °C, with ~90% and ~74% of the total viral sequences, respectively. However, cyanophage transcripts abruptly decrease at 66 °C. Cyanophages were exclusively related to the Myoviridae and Podoviridae families (Supplementary Table S3). Reads associated with cyanopodoviruses and cyanomyoviruses gradually decreased with temperature; between 48 °C to 58 °C, virus reads declined from 95% and 96% to 84% and 89%, respectively. On the other hand, at 66 °C a more severe decline was observed, to 15% and 20%, respectively. Conversely, with the reduced representation of Cyanobacteria at 66ºC, other caudovirales transcripts increased, including those that infect Proteobacteria (~31%), Firmicutes (~30%) and Actinobacteria (~23%).

Thermophilic cyanophage genome recovery.

The metagenome assembly recovered 3,912; 2,697; and 2,758 contigs, at 48°C, 58°C and 66°C, respectively. A script search (Crits-Christoph et al., 2016) resulted in 11 circular contigs, possibly indicating complete genomes. Subsequent BLASTP analysis (Camacho et al., 2009) of predicted

proteins indicated that only one circular contig had viral hallmark genes, meanwhile 9 contigs had genes associated with bacterial mobile genetic elements and 1 contig remain completely unknown. These hallmark genes are shared by many viruses but are absents from cellular genomes (Koonin et al., 2006). VirSorter tool analysis (Roux et al., 2015a) confirmed these results, obtaining the same complete putative viral contig from the 58 °C assembly, 40,740 bp long and 43.9% of GC content. This contig, TC-CHP58 (Figure 4A), was associated with a Cyanobacterial host. TC-CHP58 was present (reads recruitment) over all temperatures in Porcelana hot spring (Figure 5 and Supplementary Figure S1). At 66 ºC, TC-CHP58 was 7 fold more abundant than their putative host (measured as *Mastigocladus* RUBISCO gene abundance); at 48 ºC, the virus-host ratio was 1:1, and at 58 ºC the host was 4 fold more abundant than TC-CHP58. Metatrancriptomic reads also show that TC-CHP58 was active over all temperatures (Figure 5 and Supplementary Figure S2), but with lower transcription levels than the putative host (measured as *Mastigocladus* RUBISCO gene activity), ranging between 80-8 fold lower (Supplementary Table S4). TC-CHP58 viral DNA:RNA ratio indicated similar proportions (2.4) at 58ºC, least similar (552.9) at 66ºC; while at 48ºC the ratio was 10.4 (Supplementary Table S4).

Genomic features and organization of TC-CHP58

Complete protein prediction and annotation of TC-CHP58 using Prokka (Seemann, 2014) and BLASTP revealed 39 putative ORFs, 10 of which were viral core proteins (i.e., capsid and tail-related proteins, DNA polymerase, Terminase, etc.), 22 had no significant similarities in NCBI nr database, and 4 were present in the database but with unknown function (Table 1).

Blast analysis of the viral genes in TC-CHP58, revealed 25% to 48% identity (amino acidic level) with proteins from Cyanophage PP, PF-WMP3 and Anabaena phage A-4L, that infect freshwater filamentous Cyanobacteria such as *Phormidium*, *Plectonema* and *Anabaena* (Table 1). At the nucleotide level, there was almost no similarity to any known sequence except for a short segment of 40 nucleotides, which showed 93% similarity to a Portal protein gene sequence of *Plectonema* and P*hormidium* cyanopodoviruses (Cyanophage PP; NC_022751 and PF-WMP3; NC_009551).

Gene prediction by Prodigal indicated that the TC-CHP58 genome might be structured into two clusters, based on the transcriptional direction and putative gene functions (Figure 4A). The predicted

ORFs (Table 1) in the sense strand encode proteins involved in DNA replication and modification, such as DNA polymerase and DNA primase/helicase. Conversely, the ORFs in the antisense strand (Table 1) encode proteins necessary for virion assembly, such as major capsid protein, tail fiber proteins, internal protein/peptidase, tail tubular proteins, scaffold protein and portal protein. Moreover, two ORFs in the antisense strand had the best hits to the cyanobacterial hypothetical proteins found in the filamentous cyanobacterium *Fischerella* (WP_026731322. 1) and the unicellular *Gloeobacter* (WP_023172199.1).

Additionally, VIRFAM (Lopes et al., 2014) was used to classify TC-CHP58 according to their neck organization (Supplementary Figure S3), being assigned to the Podoviridae Type 3 category with neck structural organization similar to the Enterobacteria phage P22 (Lopes et al., 2014). Hierarchical clustering of neck proteins grouped TC-CHP58 together with the freshwater cyanophages Pf-WMP3 and Pf-WMP4, separating them from marine cyanophages such as P60 and Syn5.

Even when a large number of viral reads were assigned to cyanophages of Myoviridae family, it was not possible to recover any genome of this type. Most of the Myoviridae related contigs only had non-structural genes or hypothetical proteins of unknown function which align with proteins of known cyanomyoviruses. Here, the absence of hallmark genes from Cyanobacteria related viruses makes their accurate classification as cyanomyoviruses impossible.

Phylogenetic analysis of phage TC-CHP58.

To investigate the relationship of the phage TC-CHP58 within the Podoviridae family, the DNApol gene was selected for comparison, using published viral genomes. The analysis included representatives of Picovirinae and Autographivirinae subfamilies, plus all the available DNApol genes from known freshwater podoviruses (Pf-WMP3, PP, Pf-WMP4 and A-4L) infecting filamentous heterocystous cyanobacteria from the order Nostocales and non-heterocystous from order Oscillatoriales, plus those infecting marine *Synechococcus* spp. and *Prochlorococcus* spp. The DNApol tree (Figure 6) showed the phage TC-CHP58 as part of a monophyletic clade with all cyanopodoviruses described as infecting freshwater filamentous cyanobacteria, and more distantly, with the marine cyanopodovirus clade that infects *Synechococcus* spp. and *Prochlorococcus* spp Both cyanophage subgroups are closely related with podoviruses from the Autographivirinae subfamily, which includes all T7 relatives. Furthermore, the phylogeny of the major capsid protein (MCP) was constructed for

freshwater and marine representatives of the Autographivirinae subfamily. The available MCP gene from BHS3 Cyanophage partial genome, that is the only known thermophilic representative within the Podoviridae family, was also included (Zablocki et al., 2017). The MCP tree (Supplementary Figure S4) showed similar results to the DNApol tree (Figure 6), with a monophyletic origin for all freshwater cyanophages infecting filamentous cyanobacteria, emphasizing the division between freshwater and marine cyanobacterial viruses, and their affiliation with T7 phage. The thermophilic representatives of Podoviridae family were located in different branches inside the freshwater clade, with BHS3 more basal than TC-CHP58.

CRISPR arrays on TC-CHP58 host.

Given the high abundance (Mackenzie et al., 2013; Alcamán et al., 2015) and activity (Alcamán et al., 2015) of cyanobacteria, such as *Mastigocladus* spp., in Porcelana hot spring (Figure 3A), and in order to confirm the putative host of phage TC-CHP58, CRISPR spacer arrays were identified using the CRISPRFinder tool (Grissa et al., 2007) for seven *Mastigocladus* spp. Contigs, obtained from metagenome assemblies at 48 °C, 58 °C and 66 °C. Three CRISPR *loci* were common between all temperatures (48_CRISPR_2, 58_CRISPR_5 and 66_CRISPR_2), while four *loci* were specific to higher temperatures (58-66ºC) (Table 2). In total, the 7 CRISPR *loci* contain 562 spacers, of which 25 of them had a proto-spacer sequence in the TC-CHP58 genome (Table 2). From the 25 spacers, 19 have a target ORF of known function, such as DNA polymerase, dTMP, portal protein, M23-petidase, tail protein, tail fiber, and deoxycytidine triphosphate deaminase. In general, each CRISPR *loci* contained spacers against different ORFs on TC-CHP58, or even against different locations on the same ORF. For the 25 spacers, searching the nt/nr database, using BLASTN and BLASTX, showed no similarity to any know sequence. Finally, in order to check if CRISPR systems were active, expression of the 7 *loci* was directly quantified in the three metatrancriptomes. For all temperatures, slightly lower transcript levels were found compared to the *Mastigocladus* RUBISCO gene (Figure 5).

Identifying Single Nucleotide Variants in TC-CHP58 genome.

To assess if mismatches between the CRISPR spacer and proto-spacer sequences in TC-CHP58 genome were concealing potential variations in TC-CHP58 populations, a single nucleotide variant

(SNV) calling was conducted. For this task LoFreq tool was used, as it is high sensitivity and has low false positive rates, lower as <0.00005% (Wilm *et al*., 2012) and higher as 8.3% (Huang *et al*., 2015). This approach, together with the use of sequences with qualities over q28 (whose error probability in the base call is </= 1.58%), allow us to consider these SNVs as real mutations.

A different number of SNVs was found at each temperature. TC-CHP58 showed 1611, 930, and 671 variant sites at 48ºC, 58ºC, and 66ºC, respectively, unevenly distributed throughout the viral genome (Supplementary Figure S5). Considering the three metagenomes, a total of 3212 variable sites were present in the TC-CHP58 genome, with 391 SNVs present over all temperatures (Supplementary Figure S5). Most of the SNVs (74% on average) were located at coding regions on the TC-CHP58 genome, with variable rates, ranging from 15 to zero SNVs for each 100 pb (Supplementary Table S5) over different ORFs.

A detailed analysis of SNVs in CRISPRs proto-spacer sites revealed the presence of these polymorphisms in 14 of the 25 spacer targets, with 13 mismatches and 4 perfect matches (Table 2). The total number of polymorphic sites was 22, with 13 SNVs causing a synonymous substitution and 7 causing a non-synonymous substitution (Table 2).

DISCUSSION.

The study of viruses from thermophilic phototrophic microbial mat communities remains largely unexplored except for a few cases providing limited information on viral presence within these communities (Heidelberg et al., 2009; Davison et al., 2016). Thus far, no study has characterized viral composition and activity, or the identity of any complete viral genome. Here, using metagenomic and metatranscriptomic approaches, the composition of the most abundant and active viruses associated with the dominant members of the thermophilic bacterial community have been characterized, describing for the first time a full genome from a thermophilic cyanopodovirus (TC-CHP58). Moreover, the active cross-fire between this new cyanophage and its host is demonstrated, through TC-CHP58 population diversification (SNV), and *Mastigocladus* spp. CRISPR heterogeneity, as a response to selective pressure from the host defense system and viral predation, respectively.

Active and ubiquitous Cyanophage-type Caudovirales in phototrophic microbial mats.

The taxonomic classification of small subunit rRNA (Supplementary Table S1) indicates that the phototrophic mats in Porcelana hot spring are dominated by Bacteria (96% on average) as commonly observed in other thermophilic phototrophic microbial mats (Inskeep et al., 2013; Bolhuis et al., 2014).

Porcelana microbial mats are mainly built by filamentous representatives of two phototrophic phyla, Cyanobacteria (oxygenic) and Chloroflexi (anoxygenic), with *Mastigocladus*, *Chloroflexus* and *Roseiflexus* as the main genera, respectively. This is verified by previous surveys carried out by the authors (Mackenzie et al., 2013; Alcamán et al., 2015), as well as investigations from the White Creek, Mushroom and Octopus hot springs in Yellowstone (Miller et al., 2009; Klatt et al., 2013; Inskeep et al., 2013; Bolhuis et al., 2014), presenting similar pH, thermal gradient and low sulfide concentrations.

Porcelana dominant viruses (~70% and ~68% of metagenomic and metatranscriptomic reads) are from the families Myoviridae, Podoviridae and Siphoviridae within the Caudovirales Order (Figure 2), which typically infect Bacteria and some non-hyperthermophilic Archaea (Maniloff and Ackermann, 1998). These results were also supported by TEM images (Figure 1). The small decrease in transcripts associated to caudovirales with the increase in temperature is due to the reduction of sequences related to Podovirus and Myovirus families. A plausible explanation, is that at high temperatures some representatives of these families might have a lysogenic lifestyle, then a fraction of them will remain inactive as prophages.

Dominance by Caudovirales was only reported recently from the Brandvlei hot spring, South Africa, a slightly acidic (pH 5.7) hot spring with moderate temperature (60 ºC) and green microbial mat patches (Zablocki et al., 2017). Previously, the presence of this viral order had only been suggested in moderate thermophilic phototrophic mats from Yellowstone hot springs, through indirect genomic approximations, such as spacers in CRISPR *loci*, from dominant bacterial members (Heidelberg et al., 2009; Davison et al., 2016) or classifications based on nucleotide motives in metaviromic data (Pride and Schoenfeld, 2008; Davison et al., 2016).

Contributions from megavirus sequences were also identified in Porcelana hot spring (Figure 2), with an average of ~24% viral metagenomic reads, associated with unicellular eukaryotic hosts such as those from Phycodnaviridae and Mimiviridae families, and also the family Marseilleviridae, but to a lesser extent. The presence of VLPs from these three viral families could not be corroborated through TEM, using the limited available viral fraction (<0.2 μm) within the community, as it has been previously documented that nucleocytoplasmic large DNA viruses (NCLDV) particles are only found in larger viral fractions (Pesant et al., 2015). The ubiquity of NCLDVs in hot springs was previously described in a hydrothermal freshwater lake in Yellowstone, with assemblies of genomes from Phycodnaviridae and Mimiviridae (Zhang et al., 2015).

Viral relative abundances and activity reported here can be affected by the lack of replicates at this highly local heterogeneity samples. However, the fact of having three different temperature sampling points for metagenomics and metatranscriptomics, partially compensates the replicate limitation.
Furthermore, many viruses in an environmental sample share a degree of similarity in their genomic sequence, and this intrinsic complexity of metagenomic/metatranscriptomic samples makes difficult to accurately estimate the relative abundances or activity of specific phages at low ranks of taxonomy tree, such as the species level (Sohn *et al*., 2014). To avoid this problem, our strategy focused on the use of the LCA algorithm at higher taxonomic levels (Order and Family) to classify the viral reads, as well as for the inferred hosts, we use the phylum level.
Virus-host inference in Porcelana phototrophic mats (Figure 3B), demonstrated that the most frequent targets for viral infections were the most dominant and active components of the bacterial communities. Similarly, this is the case in other environments, such as in the human microbiome (Macklaim et al., 2013) and marine communities (Thingstad et al., 2014; Zeigler-Allen et al., 2017). In Porcelana, it is demonstrated that within microbial mats at 48 °C and 58 °C, cyanophages were among the most active viruses (Figure 3B), as were Cyanobacteria, such as *Mastigocladus* spp., as exemplified in terms of primary production and nitrogen fixation (Alcamán et al., 2015). The presence of cyanophages has been previously suggested in Yellowstone hot spring phototrophic mats (Heidelberg et al., 2009; Davison et al., 2016), and more recently in the Brandvlei hot spring, South Africa (Zablocki et al., 2017). Heidelberg et al, (2009) found that CRISPR spacers in unicellular cyanobacteria *Synechococcus* isolates (Syn OS-A and Syn OS-B9) from Octopus Hot Spring, might have 23 known viral targets (lysozyme-related reads, PFAM DUF847) on an independently published metavirome from the same

hot spring. More recently, 171 viral contigs associated with the host genus S*ynechococcus*, based on tetranucleotide frequencies, were identified from a microbial mat (60°C) metavirome from Octopus Spring. The majority of the annotated ORFs on the viral contigs coded for glycoside hydrolases, with lysozyme activity, identifying 6 CRISPR proto-spacers in those genes (Davison et al., 2016). Even though a taxonomic relationship with cyanophages was not confirmed for those proto-spacers containing contigs (Heidelberg et al., 2009; Davison et al., 2016), it provides evidence towards the presence of cyanophages related sequences within these thermophilic mats. The work by Zablocki et al, (2009) reconstructed a 10 kb partial genome of a new cyanophage (BHS3) from Brandvlei hot spring metavirome, stating that cyanophages appear to be the dominant viruses in the hot spring. The BHS3 contig (MF098555) contains 9 ORFs, with the majority of the identified proteins having a close relation to the Cyanophage PP and *Phormidium* phage Pf-WMP3, which infect freshwater filamentous cyanobacteria *Phormidium* and *Plectonema*.

The presence of cyanophages related sequences in thermophilic phototrophic mats is significant, since these viruses are known to play an important role in the evolution of cyanobacteria (Shestakova and Karbysheva, 2015). Cyanophages affect the rate and direction of cyanobacterial evolutionary processes, through the regulation of abundance, population dynamics, and natural community structure. This has been extensively studied and demonstrated for marine environments (Weinbauer and Rassoulzadegan, 2004; Avrani et al., 2011). These cyanophages are proven to play a relevant role in the marine biogeochemical cycles, through the infection and lysis of Cyanobacteria, affecting carbon and nitrogen fixation (Suttle, 2000). Moreover, cyanophages act as a global reservoir of genetic information, as they are vectors for gene transfer, meaning that cyanobacteria can acquire novel attributes within aquatic environments (Kristensen et al., 2010; Chénard et al., 2018).

Caudoviruses were prevalent at 66°C in Porcelana, and potentially infecting Firmicutes, Proteobacteria and Actinobacteria. These phila have also been previously identified in other hot springs at temperatures above 76 °C, such as in Octopus and Bear Paw (Pride and Schoenfeld, 2008). At high temperatures in Porcelana also the phylum Chloroflexi was dominant in the phototrophic mat (Figure 3A). However, viral sequences related to this taxon could not be retrieved, as neither viruses nor viral sequences have been confirmed to infect members of this phylum in any environment. Davison et al, (2016), described viral contigs associated with *Roseiflexus* sp. from a metavirome from Octopus

Spring, but only raw reads are publicly available, without taxonomic assignation. Finally, the recently released IMG/VR database (Paez-Pino et al., 2016) contains 3 contigs associated by CRISPR spacers to *Chloroflexus* sp. Here, a BLASTP analyses against RefSeq viral proteins revealed that 6 of these proteins have a best hit in *Mycobacterium* phage proteins and one which best hit was a *Clavibacter* phage protein. These findings, suggest that some of the viral reads classified as Actinobacteria viruses could be instead from unknown Chloroflexi viruses.

Viral mining reveals a new infective thermophilic cyanopodovirus lineage.

Metagenomic surveys of viral genomes are an effective way to detect unknown viruses (Voorhies et al., 2015; Zhang et al., 2015; Roux et al., 2015a, 2015b). In metagenomics, two key elements for virus detection are the presence of viral hallmark genes and the circularity of viral contigs (Roux et al., 2015a, 2015b). Based on these two principles, a complete genome (TC-CHP58) was identified. The genome was represented by a viral contig of 50 kb, which is a typical size for Caudovirales members from the Podoviridae family. The genome size and viral core proteins affiliated with the Podovirus seems to make TC-CHP58 the first report of a full genome of a thermophilic cyanopodovirus. Moreover, the genome organization (Figure 4B) shows a consistent synteny with other cyanopodoviruses, which also lack RNA polymerase inside the T7 supergroup, as described for the viruses Pf-WMP4, Pf-WMP3, Cyanophage PP, Anabaena phage A-4L (Liu et al., 2007, 2008; Zhou et al., 2013; Ou et al., 2015), and the recently reported partial genome of the thermophilic BHS3 cyanophage (Zablocki et al., 2017). Initially, the presence of a single-subunit RNA polymerase that binds phage specific promoters was considered to be a major, and unique characteristic of the T7 supergroup (Dunn et al., 1983). However, more recently, it has been proposed that podoviruses that share extensive homology with T7, but lack the phage RNA polymerase, are still part of the T7 supergroup, as distant and probably ancient branches (Hardies et al., 2003).

TC-CHP58 presented a genome organization that can be divided into two portions (Figure 4A); with ORFs in the sense strand related to DNA replication and modification, and genes encoded in the antisense strand related to virion assembly. This genome organization is also present in other freshwater T7-related podoviruses that infect filamentous cyanobacteria (Liu et al., 2007, 2008; Zhou et al., 2013; Ou et al., 2015), including the thermophilic BHS3 cyanophage (Zablocki et al., 2017). This setup is

548 also similar to the class II and III organization genes in T7-like viruses, where class II genes are
549 responsible for DNA replication and metabolism, and class III genes include structural and maturation
550 genes (Dunn et al., 1983). The VIRFAM analysis of neck protein organization verifies the classification
551 of TC-CHP58 within the Podoviridae family (Supplementary Figure S3), where the Type 3 podovirus
552 encompasses T7-like phages from Autographinavirinae subfamilies and several other genera (Lopes et
553 al., 2014). The T7-like classification for TC-CHP58, and other podoviruses that infect freshwater
554 filamentous cyanobacteria, is supported by the organization of the genome into two portions as well as
555 the organization of the neck proteins.
556
557 The phylogenetic position of TC-CHP58, based on DNA polymerase I (DNApol) (Figure 6) and Major
558 capsid (MCP) (Supplementary Figure S4) predicted proteins, confirm the affiliation of this new virus
559 within the family Podoviridae. Both phylogenetic markers verify the separation between the marine
560 from the freshwater cyanopodoviruses within the T7 family, as previously proposed (Liu et al., 2007;
561 Ou et al., 2015). These results also support the connection between the T7 phages and marine and
562 freshwater cyanopodoviruses (Chen and Lu, 2002; Hardies et al., 2003; Liu et al., 2007; Ou et al.,
563 2015), including TC-CHP58 and BHS3 as representatives of a novel, and potentially globally
564 distributed thermophilic cyanophage lineage. Moreover, this data demonstrates that marine and
565 freshwater cyanopodoviruses, including the thermophilic TC-CHP58, are part of the Autographivirinae
566 subfamily as previously suggested for Cyanophage P60 and Roseophage SIO1 (Labonté et al., 2009),
567 both included in this analysis.
568
569 In Porcelana, the virus host ratio relating to TC-CHP58 presence was lower than the typical values
570 observed in freshwater environments (Maranger and Bird, 1995), being more similar to other
571 geothermal environments where viral density is typically lower, with 10–100-fold less viruses than host
572 cells (López-López et al., 2013). This is expected, considering that there are abundant cyanobacteria in
573 phototrophic mats in Porcelana in comparison with the 10^4 mL^{-1} VLPs observed in the water of hot
574 springs (Breitbart et al., 2004). It is also demonstrated that TC-CHP58 presented higher infection
575 efficiency, as revealed by the viral DNA to RNA ratios at lower temperatures (58ºC, then 48ºC) with
576 cyanobacteria dominating, while at 66ºC most of the TC-CHP58 remained inactive (Figure 5).
577 Infection inefficiency is multidimensional, as it initiates from reduced phage adsorption, RNA, DNA
578 and protein production (Howard-Varona et al., 2017). Thus, the high copy number of TC-CHP58 DNA

579 at 66ºC may be due to the persistence of viral DNA (Mengoni et al., 2005) encapsidated extracellularly
580 and intermixed in the microbial mat were the host (*Mastigocladus* spp) has a low activity as evidenced
581 by the low expression of the RUBISCO gene and the CRISPR *loci*. An alternative explanation is the
582 absence, or the diminished presence, of the specific host due to intraspecific diversification as
583 evidenced by the existence of different CRISPR *loci* at different temperatures. This theory has been
584 proposed for other cyanobacteria, such as *Prochlorococcus* and *Phormidium*, where slight differences
585 in fitness, niche, and selective phage predation, explain the coexistence of different populations
586 (Kashtan et al., 2014; Voorhies et al., 2016). The last explanation acquires special importance in light of
587 recent evidence that variations in the structure and function of the heterocyst and differential CRISPR
588 *loci* are fundamental to diversification of *Mastigocladus laminosus* (also known as *Fischerella*
589 *thermalis*), a cosmopolitan thermophilic cyanobacterium, reinforcing the importance of viral predation
590 (Sano et al., 2018).
591
592 **CRISPR spacers assign *Mastigocladus* spp. as putative hosts for TC-CHP58.**
593
594 It was possible to verify *Mastigocladus* spp. as putative hosts for the new cyanopodovirus (TC-
595 CHP58), via the analysis of CRISPR spacers found in the cyanobacteria, recovered from contigs
596 obtained in the same metagenomic datasets. This methodology has been previously used for the
597 identification of novel viruses in hot springs (Heidelberg et al., 2009; Snyder et al., 2010; Davison et
598 al., 2016), as well as in other environments such as acid mines (Andersson and Banfield, 2008), the
599 human microbiome (Stern et al., 2012), as well as sea ice and soils (Sanguino et al., 2015).
600
601 Observations from the CRISPR *loci* over all temperatures (Table 2) indicated that, in general, proto-
602 spacers in the TC-CHP58 genome were distributed on coding, and therefore more conserved regions.
603 The expression of 7 CRISPR *loci* (Figure 5), demonstrated the activity of the *Mastigocladus* spp.
604 defense system against TC-CHP58 over all temperatures. CRISPR arrays are transcribed into a long
605 precursor, containing spacers and repeats, that are processed into small CRISPR RNAs (crRNAs) by
606 dedicated CRISPR-associated (Cas) endoribonucleases (Brouns et al., 2008). Although it is not possible
607 to measure mature crRNAs, as due to their small size they are likely to be filtered out in RNA-seq
608 libraries, this approximation has been validated using large datasets (Ye and Zhang, 2016).

Despite variations in the number of CRISPR *loci* observed at each temperature, with 60% of the total CRISPR *loci* found in Mastigocladus contigs at 58 ºC, the abundance of reads agreed with the abundance of other genes required by these cyanobacteria, such as the RUBISCO gene (Figure 5). This further verified that the *loci* are from *Mastigocladus* populations. The different CRISPR *loci* found over the different temperatures in Porcelana (Table 2), also reinforces the notion that diversification of *Mastigocladus* is partly due to selective pressure exerted by the predation of viruses, such as TC-CHP58. This theory has been previously put forward for *Mastigocladus laminosus* in Yellowstone (Sano et al., 2018), and proposed for marine cyanobacteria (Rodriguez-Valera et al., 2009; Kashtan et al., 2014).

Furthermore, each CRISPR *loci* contains spacers that corresponds to different proto-spacers in the TC-CHP58 genome. Increases in spacer number and diversity against the same virus may explain the increase in interference, whilst decreasing the selection of escape mutants (Staals et al., 2016). Priming mechanisms are the most efficient form of obtaining new spacers (Staals et al., 2016), using a partial match between a pre-existing spacer and the genome of an invading phage to rapidly acquire a new "primed" spacer (Westra et al., 2016). Then, over-representation of spacer sequences in some regions of the TC-CHP58 genome may be related to a site that has already been sampled by the CRISPR–Cas machinery or by other biases such as the secondary structure of phage ssDNA, GC content, and transcriptional patterns (Paez-Espino et al., 2012).

The selection pressure of multiple spacers in *Mastigocladus* CRISPR *loci* leads to the emergence of single nucleotide variants (SNVs) in the TC-CHP58 viral populations (Table 2), which cause mismatches between spacers and proto-spacers, resulting in the attenuation or evasion of the host immune response (Shmakov et al., 2017). It is still possible to utilize mismatched spacers for interference and/or primed adaptation, however the degree of tolerance to mismatches for interference among the CRISPR-Cas, varies substantially between different CRISPR-Cas type systems (Shmakov et al., 2017). The variable frequency (0.6 to 0.02) of the corresponding spacer SNVs alleles on TC-CHP58 proto-spacers, suggests that some variants are more prevalent throughout the population, regardless of whether the SNV causes a silent mutation. Based on this evidence, it has been proposed that, for other microbial communities, only the most recently acquired spacer can exactly match the

639 virus. This suggests that community stability is driven by compensatory shifts in host resistance levels
640 and virus population structure (Andersson and Banfield, 2008).
641
642 The present study describes the underlying viral community structure and activity of thermophilic
643 phototrophic mats. Moreover, abundant virus populations are linked to dominant bacteria,
644 demonstrating the effectiveness of omics approaches in estimating the importance and activity of a
645 viral community, in this case with thermophilic cyanophages.
646
647 Additionally, the first full genome of a new T7-related virus that infects thermophilic representatives of
648 the cyanobacterium *Mastigocladus* spp. was here retrieved. This genome may represent a novel,
649 globally present, freshwater thermophilic virus from a new lineage from the Podoviridae family. The
650 latter was strongly suggested by the significant phylogenetic relationship and shared gene organization
651 with the BHS3 cyanophage partial genome (South Africa). Even more, TC-CHP58 proteins also
652 matches several contigs that include common viral hallmarks genes in the IMG/VR database. However,
653 further work is necessary to fully understand the global representation and relevance of this virus,
654 which complete genome is presented here as first reference available..
655
656 Finally, the evolutionary arms race between a specific cyanobacteria-cyanophage in the natural
657 environment is exposed, where there exist a variety of potential scenarios. For instance, host resistance
658 may increase over time forcing the decrease of viral populations, or a specific virus population may
659 occasionally become extremely virulent and cause the crash of the host population as proposed by the
660 “kill the winner” model (Andersson and Banfield, 2008). Alternatively, if CRISPR systems and the
661 diversification of the viral population remain in balance through time, a relatively stable virus and host
662 community may result.
663

664 **CONFLICT OF INTEREST**

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

AUTHOR CONTRIBUTIONS

SGL and BD conceived and designed the experiments. SGL, OS and FP performed the experiments. SGL, CPA, OS, FP and BD analyzed the data. SGL, CPA, and BD wrote the paper.

FUNDING

This work was financially supported by PhD scholarships CONICYT N° 21130667, 21172022 and CONICYT grant FONDECYT N°1150171. Sequencing was funded by Spanish grant CTM2013-48292-C3-1-R.

ACKNOWLEDGMENTS

We are grateful to Huinay Scientific Field Station for making our work in the Porcelana hot spring possible.

REFERENCES

Aguilera Á, Souza-Egipsy V, González-Toril E, Rendueles O, Amils R. (2010). Eukaryotic microbial diversity of phototrophic microbial mats in two Icelandic geothermal hot springs. Int Microbiol 13: 21–32.

Amaral-Zettler LA. (2012). Eukaryotic diversity at pH extremes. Front Microbiol 3: 1–17.

Andersson AF, Banfield JF. (2008). Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science (80-) 320: 1047–1050.

Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. (2011). Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature 474: 604–8.

Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19: 455–477.

Bhaya, D., Grossman, A. R., Steunou, A.-S., Khuri, N., Cohan, F. M., Hamamura, N., et al. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. 1, 703–713.

Bize A, Peng X, Prokofeva M, Maclellan K, Lucas S, Forterre P, et al. (2008). Viruses in acidic geothermal environments of the Kamchatka Peninsula. Res Microbiol 159: 358–66.

Bolduc B, Wirth JF, Mazurie A, Young MJ. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. ISME J 9: 1–16.

Bolhuis H, Cretoiu MS, Stal LJ. (2014). Molecular ecology of microbial mats. FEMS Microbiol Ecol 90: 335–350.

Breitbart M, Wegley L, Leeds S, Rohwer F, Schoenfeld T. (2004). Phage Community Dynamics in Hot Springs. Appl Environ Microbiol 70: 1633–1640.

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science (80-.). 321, 960 LP-964.

Brown PB, Wolfe G V. (2006). Protist genetic diversity in the acidic hydrothermal environments of Lassen Volcanic National Park, USA. J Eukaryot Microbiol 53: 420–431.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. (2009). BLAST plus: architecture and applications. BMC Bioinformatics 10: 1.

Chen F, Lu J. (2002). Genomic Sequence and Evolution of Marine Cyanophage P60 : a New Insight on Lytic and Lysogenic Phages Genomic Sequence and Evolution of Marine Cyanophage P60 : a New Insight on Lytic and Lysogenic Phages . Appl Environ Microbiol 68: 2589–2594.

Cole JK, Peacock JP, Dodsworth JA, Williams AJ, Thompson DB, Dong HL, et al. (2013). Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. Isme J 7: 718–729.

Crits-Christoph A, Gelsinger DR, Ma B, Wierzchos J, Ravel J, Davila A, et al. (2016). Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. Environ Microbiol 18: 2064–2077.

Cuervo A, Pulido-Cid M, Chagoyen M, Arranz R, González-García VA, Garcia-Doval C, et al. (2013). Structural characterization of the bacteriophage T7 tail machinery. J Biol Chem 288: 26290–26299.

Darriba D, Taboada GL, Posada D. (2011). ProtTest 3 : fast selection of best-fit models of protein evolution. Bioinformatics 27: 1164–1165.

Davison M, Treangen TJ, Koren S, Pop M, Bhaya D. (2016). Diversity in a Polymicrobial Community Revealed by Analysis of Viromes, Endolysins and CRISPR Spacers. PLoS One 11: e0160574.

Diemer GS, Stedman KM. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA virus. Biol Direct 7: 1–14.

Dunn JJ, Studier FW, Gottesman M. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 166: 477–535.

Edgar RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.

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

Goren MG, Yosef I, Qimron U. (2015). Programming Bacteriophages by Swapping Their Specificity Determinants. Trends Microbiol 23: 744–746.

757 Grissa I, Vergnaud G, Pourcel C. (2007). CRISPRFinder: a web tool to identify clustered regularly
758 interspaced short palindromic repeats. Nucleic Acids Res 35: 52–57.
759
760 Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. (2016). Microbial Community Analysis with
761 Ribosomal Gene Fragments from Shotgun Metagenomes. Appl Environ Microbiol 82: 157–166.
762
763 Ha M, Rachel R, Peng X, Garrett RA, Prangishvili D. (2005). Viral Diversity in Hot Springs of
764 Pozzuoli, Italy, and Characterization of a Unique Archaeal Virus,. J Virol 79: 9904–9911.
765
766 Hardies SC, Comeau AM, Serwer P, Suttle CA. (2003). The complete sequence of marine
767 bacteriophage VpV262 infecting Vibrio parahaemolyticus indicates that an ancestral component of a T7
768 viral supergroup is widespread in the marine environment. Virology 310: 359–371.
769
770 Hargreaves KR, Flores C, Lawley T, Cloki M. (2014). Abundant and Diverse Clustered Regularly
771 Interspaced Short Palindromic Repeat Spacers in Clostridium difficile Strains and Prophages Target
772 Multiple Phage Types within This Pathogen Katherine. MBio 5: 1–10.
773
774 Heidelberg JF, Nelson WC, Schoenfeld T, Bhaya D. (2009). Germ Warfare in a Microbial Mat
775 Community: CRISPRs Provide Insights into the Co-Evolution of Host and Viral Genomes. PLoS One
776 4: e4169.
777
778 Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. (2016). MEGAN Community
779 Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS
780 Comput Biol 12: 4–12.
781
782 Huang HW, Mullikin JC, Hansen NF. (2015). Evaluation of Variant Detection Software for Pooled
783 Next-Generation Sequence Data. BMC Bioinformatics 16:235.
784
785 Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. (2010). Prodigal: prokaryotic
786 gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.
787

Inskeep WP, Jay ZJ, Tringe SG, Herrgård MJ, Rusch DB. (2013). The YNP Metagenome Project: Environmental Parameters Responsible for Microbial Distribution in the Yellowstone Geothermal Ecosystem. Front Microbiol 4: 67.

Jonker, C.Z.; van Ginkel, C.; Olivier, J. Association between physical and geochemical characteristics of thermal springs and algal diversity in Limpopo Province, South Africa. Water SA 2013, 39, 95–104.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science (80-.). 344, 416–420.

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.

Klatt CG, Inskeep WP, Herrgard MJ, Jay ZJ, Rusch DB, Tringe SG, et al. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. Front Microbiol 4: 1–23.

Klatt CG, Wood JM, Rusch DB, Bateson MM, Hamamura N, Heidelberg JF, et al. (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. ISME J 5: 1262–78.

Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. ISME J. 11, 2399–2406.

Koonin E V, Senkevich TG, Dolja V V. (2006). The ancient Virus World and evolution of cells. Biol Direct 1: 27.

Kristensen DM, Mushegian AR, Dolja V V, Koonin E V. (2010). New dimensions of the virus world discovered through metagenomics. Trends Microbiol 18: 11–9.

Kumar JK, Chiu ET, Tabor S, Richardson CC. (2004). A unique region in bacteriophage T7 DNA polymerase important for exonucleolytic hydrolysis of DNA. J Biol Chem 279: 42018–42025.

Labonté JM, Reid KE, Suttle CA, Labont JM, Reid KE, Suttle CA. (2009). Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. Appl Environ Microbiol 75: 3634–3640.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359.

Liu X, Kong S, Shi M, Fu L, Gao Y, An C. (2008). Genomic analysis of freshwater cyanophage Pf-WMP3 infecting cyanobacterium Phormidium foveolarum: The conserved elements for a phage. Microb Ecol 56: 671–680.

Liu X, Shi M, Kong S, Gao Y, An C. (2007). Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium Phormidium foveolarum: Complete genome sequence and DNA translocation. Virology 366: 28–39.

Liu, Z., Klatt, C. G., Wood, J. M., Rusch, D. B., Ludwig, M., Wittekindt, N., et al. (2011). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. ISME J. 5, 1279–1290.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol 16: 2659–2671.

Lopes, A., Tavares, P., Petit, M. A., Guérois, R., and Zinn-Justin, S. (2014). Automated classification of tailed bacteriophages according to their neck organization. BMC Genomics 15, 1–17.

López-López, O., Cerdán, M., and González-Siso, M. (2013). Hot Spring Metagenomics. Life 3, 308–320.

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: 123–36.

Macklaim JM, Fernandes AD, Di Bella JM, Hammond J-A, Reid G, Gloor GB. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome 1: 12.

Maniloff J, Ackermann HW. (1998). Taxonomy of bacterial viruses: Establishment of tailed virus genera and the order Caudovirales. Arch Virol 143: 2051–2063.

Maranger, R., and Bird, D. F. (1995). Viral abundance in aquatic systems: A comparison between marine and fresh waters. Mar. Ecol. Prog. Ser. 121, 217–226.

Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10.

Mihara T, Nishimura Y, Shimizu Y, Nishiyama H, Yoshikawa G, Uehara H, et al. (2016). Linking virus genomes with host taxonomy. Viruses 8: 10–15.

Miller SR, Purugganan M, Curtis SE. (2006). Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium Mastigocladus laminosus. Appl Environ Microbiol 72:2793-2800.

Miller SR, Strong AL, Jones KL, Ungerer MC. (2009). Bar-coded pyrosequencing reveals shared bacterial community properties along the temperature gradients of two alkaline hot springs in Yellowstone National Park. Appl Environ Microbiol 75: 4565–4572.

Notarnicola SM, Mulcahy HL, Lee J, Richardson CC. (1997). The Acidic Carboxyl Terminus of the Bacteriophage T7 Gene 4 Helicase / Primase Interacts with T7 DNA Polymerase. J Biol Chem 272: 18425–18433.

880 Ou T, Liao XY, Gao XC, Xu XD, Zhang QY. (2015). Unraveling the genome structure of
881 cyanobacterial podovirus A-4L with long direct terminal repeats. Virus Res 203: 4–9.
882
883 Paez-Espino, D., Morovic, W., Sun, C. L., Thomas, B. C., Ueda, K. I., Stahl, B., et al. (2013). Strong
884 bias in the bacterial CRISPR elements that confer immunity to phage. Nat. Commun. 4, 1430–1437.
885
886 Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova,
887 N., et al. (2016). Uncovering Earth's virome. Nature 536, 425–430.
888
889 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM:
890 assessing the quality of microbial genomes recovered from. Cold Spring Harb. Lab. Press Method 1, 1–
891 31.
892
893 Pawlowski A, Rissanen I, Bamford JKH, Krupovic M, Jalasvuori M. (2014). Gammasphaerolipovirus,
894 a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria
895 within the novel family Sphaerolipoviridae. Arch Virol 159: 1541–1554.
896 Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, et al. (2015). Open science
897 resources for the discovery and analysis of Tara Oceans data. Sci data 2: 150023.
898
899 Prangishvili D, Garrett R a. (2004). Exceptionally diverse morphotypes and genomes of crenarchaeal
900 hyperthermophilic viruses. Biochem Soc Trans 32: 204–8.
901
902 Pride DT, Schoenfeld T. (2008). Genome signature analysis of thermal virus metagenomes reveals
903 Archaea and thermophilic signatures. BMC Genomics 9: 420.
904
905 Quinlan, A. R., and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic
906 features. Bioinformatics 26, 841–842.
907
908 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. (2013). The SILVA ribosomal RNA
909 gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41: 590–596.
910

911 Redder P, Peng X, Brügger K, Shah S a, Roesch F, Greve B, et al. (2009). Four newly isolated
912 fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible
913 interviral recombination mechanism. Environ Microbiol 11: 2849–62.
914
915 Richter M, Rosselló-Móra R, Glöckner F, Peplies J. (2016) JSpeciesWS: a web server for prokaryotic
916 species circumscription based on pairwise genome comparison, Bioinformatics, Volume 32, (6):929–
917 931.
918
919 Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasić, L., Thingstad, T. F., Rohwer,
920 F., et al. (2009). Explaining microbial population genomics through phage predation. Nat. Rev.
921 Microbiol. 7, 828–36.
922
923 Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. (2012). Mrbayes 3.2:
924 Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:
925 539–542.
926
927 Roux S, Enault F, Hurwitz BL, Sullivan MB. (2015a). VirSorter: mining viral signal from microbial
928 genomic data. PeerJ 3: e985.
929
930 Roux S, Hallam SJ, Woyke T, Sullivan MB. (2015b). Viral dark matter and virus-host interactions
931 resolved from publicly available microbial genomes. Elife 4: e08490.
932
933 Sanguino L, Franqueville L, Vogel TM, Larose C. (2015). Linking environmental prokaryotic viruses
934 and their host through CRISPRs. FEMS Microbiol Ecol 91: 1–9.
935
936 Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. (2018). Ancient balancing selection on
937 heterocyst function in a cosmopolitan cyanobacterium. Nat. Ecol. Evol., 1–10.
938
939 Schmieder R, Edwards R. (2011). Quality control and preprocessing of metagenomic datasets.
940 Bioinformatics 27: 863–864.
941

942 Schmieder R, Lim YW, Edwards R. (2012). Identification and removal of ribosomal RNA sequences
943 from metatranscriptomes. Bioinformatics 28: 433–435.
944
945 Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D. (2008). Assembly of
946 Viral Metagenomes from Yellowstone Hot Springs. Appl Environ Microbiol 74: 4164–4174.
947
948 Seemann T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics 30: 2068–2069.
949 Shestakova S V., Karbysheva EA. (2015). The role of viruses in the evolution of Cyanobacteria. Biol
950 Bull Rev 5: 527–537.
951
952 Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, K. E. (2017). The CRISPR Spacer
953 Space Is Dominated by crossm The CRISPR Spacer Space Is Dominated. MBio 8, 1–18.
954
955 Snyder JC, Bateson MM, Lavin M, Young MJ. (2010). Use of cellular CRISPR (clusters of regularly
956 interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in
957 environmental samples. Appl Environ Microbiol 76: 7251–8.
958
959 Sohn MB, An L, Pookhao N, Li Q. (2014). Accurate genome relative abundance estimation for closely
960 related species in a metagenomic sample. BMC Bioinformatics 15(1).
961
962 Stern A, Mick E, Tirosh I, Sagy O, Sorek R. (2012). CRISPR targeting reveals a reservoir of common
963 phages associated with the human gut microbiome. Genome Res 22: 1985–1994.
964
965 Steunou, A., Bhaya, D., Bateson, M. M., Melendrez, M. C., Ward, D. M., Brecht, E., et al. (2006). In
966 situ analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria
967 inhabiting hot spring microbial mats. Proc. Natl. Acad. Sci. 103, 2398-2403.
968
969 Stewart W. (1970). Nitrogen fixation by blue-green algae in Yellowstone thermal areas. Phycologia
970 9:261-268.
971

Tekere, M.; Lötter, A.; Olivier, J.; Jonker, N.; Venter, S. (2011) Metagenomic analysis of bacterial diversity of Siloam hot water spring, Limpopo, South Africa. Afr. J. Biotechnol 10:18005-18012.

Thingstad, T. F., Vage, S., Storesund, J. E., Sandaa, R.-A., and Giske, J. (2014). A theoretical analysis of how strain-specific viruses can control microbial species diversity. Proc. Natl. Acad. Sci. 111, 7813–7818.

Trifinopoulos J, Nguyen L-T, von Haeseler A, Minh BQ. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44: 1–4.

Uldahl Kristine and Peng Xu. (2013). Biology, Biodiversity and Application of Thermophilic Viruses. In: Satyanarayana Tulasi, Littlechild Jennifer KY (ed). Thermophilic Microbes in Environmental and Industrial Biotechnology. Pp 271–306.

Van der Meer MT, Klatt CG, Wood J, Bryant DA, Bateson MA, Lammerts L, et al. (2010). Cultivation and Genomic, Nutritional, and Lipid Biomarker Characterization of Roseiflexus Strains Closely Related to Predominant In Situ Populations Inhabiting Yellowstone Hot Spring Microbial Mats. J Bacteriol 12: 3033–3042.

Vollmer W, Joris B, Charlier P, Foster S. (2008). Bacterial peptidoglycan (murein) hydrolases. FEMS Microbiol Rev 32: 259–286.

Voorhies A a., Eisenlord SD, Marcus DN, Duhaime MB, Biddanda BA, Cavalcoli JD, et al. (2015). Ecological and genetic interactions between cyanobacteria and viruses in a low-oxygen mat community inferred through metagenomics and metatranscriptomics. Environ Microbiol 18: 358–371.

Weinbauer MG, Rassoulzadegan F. (2004). Are viruses driving microbial diversification and diversity? Environ Microbiol 6: 1–11.

Westra ER, Dowling AJ, Broniewski JM, van Houte S. (2016). Evolution and Ecology of CRISPR. Annu Rev Ecol Evol Syst 47: 307–331.

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201.

Wu YW, Simmons BA, and Singer SW, "MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets", Bioinformatics, 32(4): 605-607, 2016.

Xie C, Goi CLW, Huson D, Little P, Williams R. (2016). RiboTagger: fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or metatranscriptome surveys. BMC Bioinformatics 17: 277–295.

Ye, Y., and Zhang, Q. (2016). Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. Rna 22, 945–956.

Zablocki, O., van Zyl, L. J., Kirby, B., and Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. Viruses 9 (11), 348.

Zeigler-Allen L, McCrow JP, Ininbergs K, Dupont CL, Badger JH, Hoffman JM, et al. (2017). The Baltic Sea Virome: Diversity and Transcriptional Activity of DNA and RNA Viruses. mSystems 2: e00125-16.

Zhang W, Zhou J, Wang Y. (2015). Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. Sci Rep 5: 15131.

Zhou J, Sun D, Childers A, McDermott TR, Wang Y, Liles MR. (2015). Three novel virophage genomes discovered from yellowstone lake metagenomes. J Virol 89: 1278–85.

Zhou Y, Lin J, Li N, Hu Z, Deng F. (2013). Characterization and genomic analysis of a plaque purified strain of cyanophage PP. Virol Sin 28: 272–279.

Data Availability Statement

The datasets generated for this study can be found NCBI as follow: Access to raw data for metagenomes and metatranscriptomes is available through NCBI BioProject ID PRJNA382437. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA382437

The genome of TC-CHP58 has the GenBank accession number KY888885. Contigs containing CRISPRs *loci* have been submitted to NCBI with GenBank accession numbers MG734911 to MG734917.

Figures

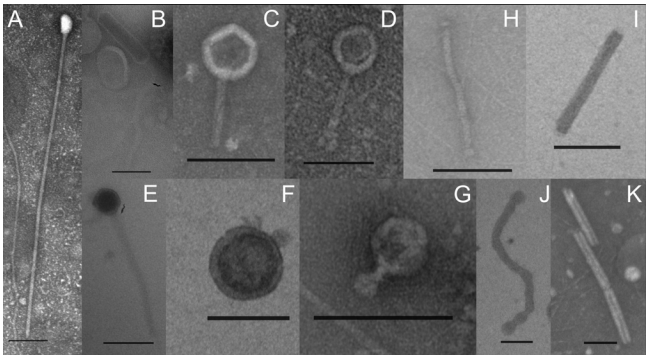


Figure 1. Transmission electronic micrographs of VLPs obtained from the interstitial fluid of phototrophic microbial mats growing between 62 °C and 42 °C in Porcelana hot spring. Scale bar: 100 nm.

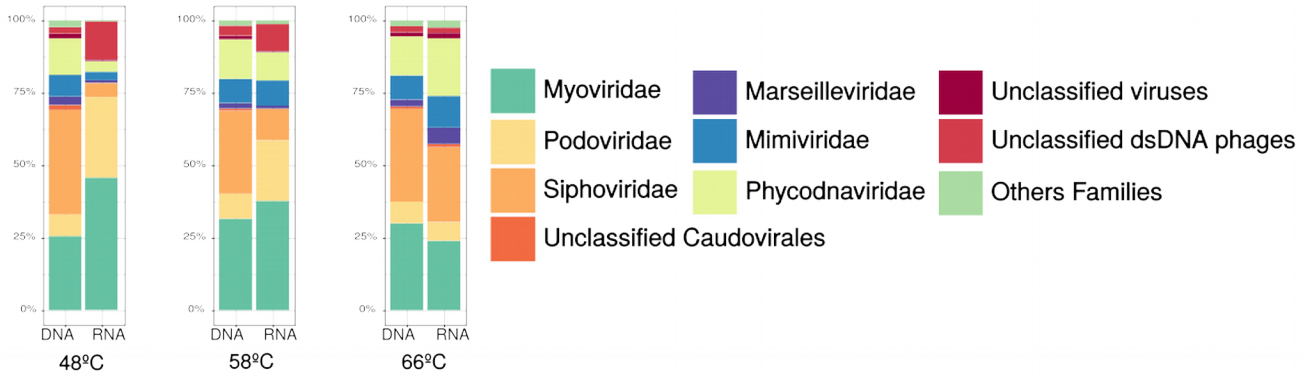
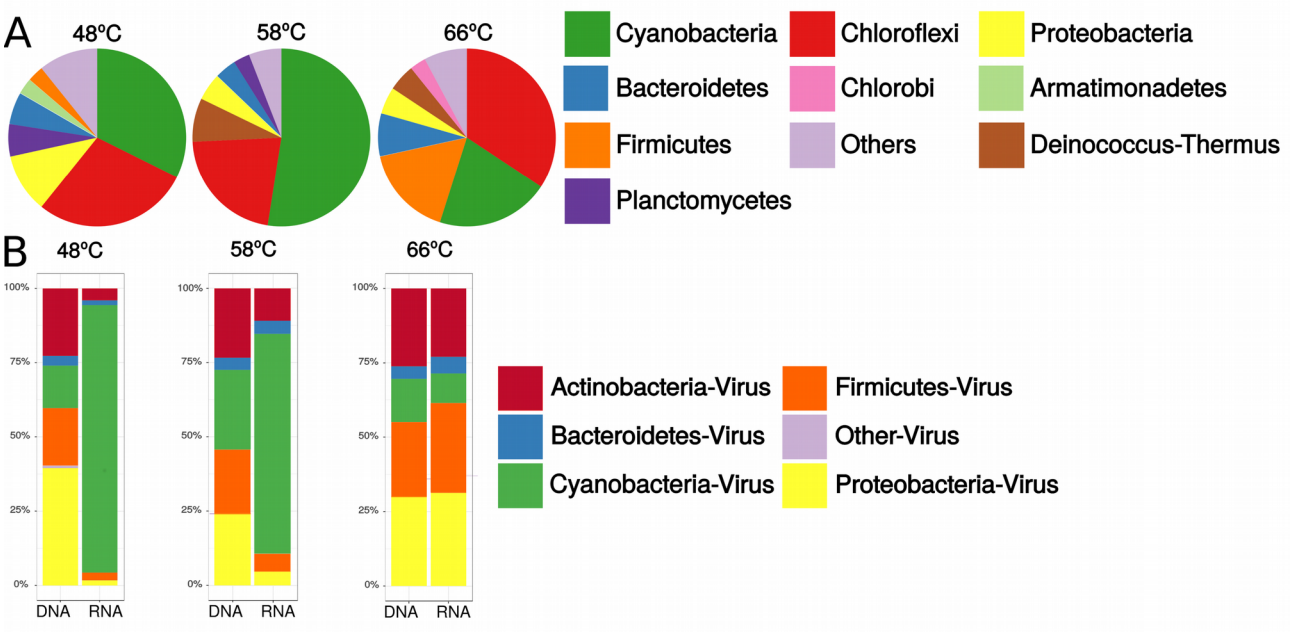


Figure 2. Relative abundances of viral Families in microbial mats from Porcelana hot spring; standardized by the total number metagenomic reads (DNA), and metatranscriptome (RNA) from each temperature samples.



1102 **Figure 3.** Relative abundances of A) Bacterial community, to Phylum level, in the microbial mats
1103 obtained from 16S miTAGs, standardized by the total number of metagenomic reads (DNA) for each
1104 temperature sample, and B) Caudovirales community at the host Phylum level, obtained from shotgun
1105 sequences in metagenomes (DNA) and metatranscriptomes (RNA), standardized by the total number of
1106 reads from each temperature sample.

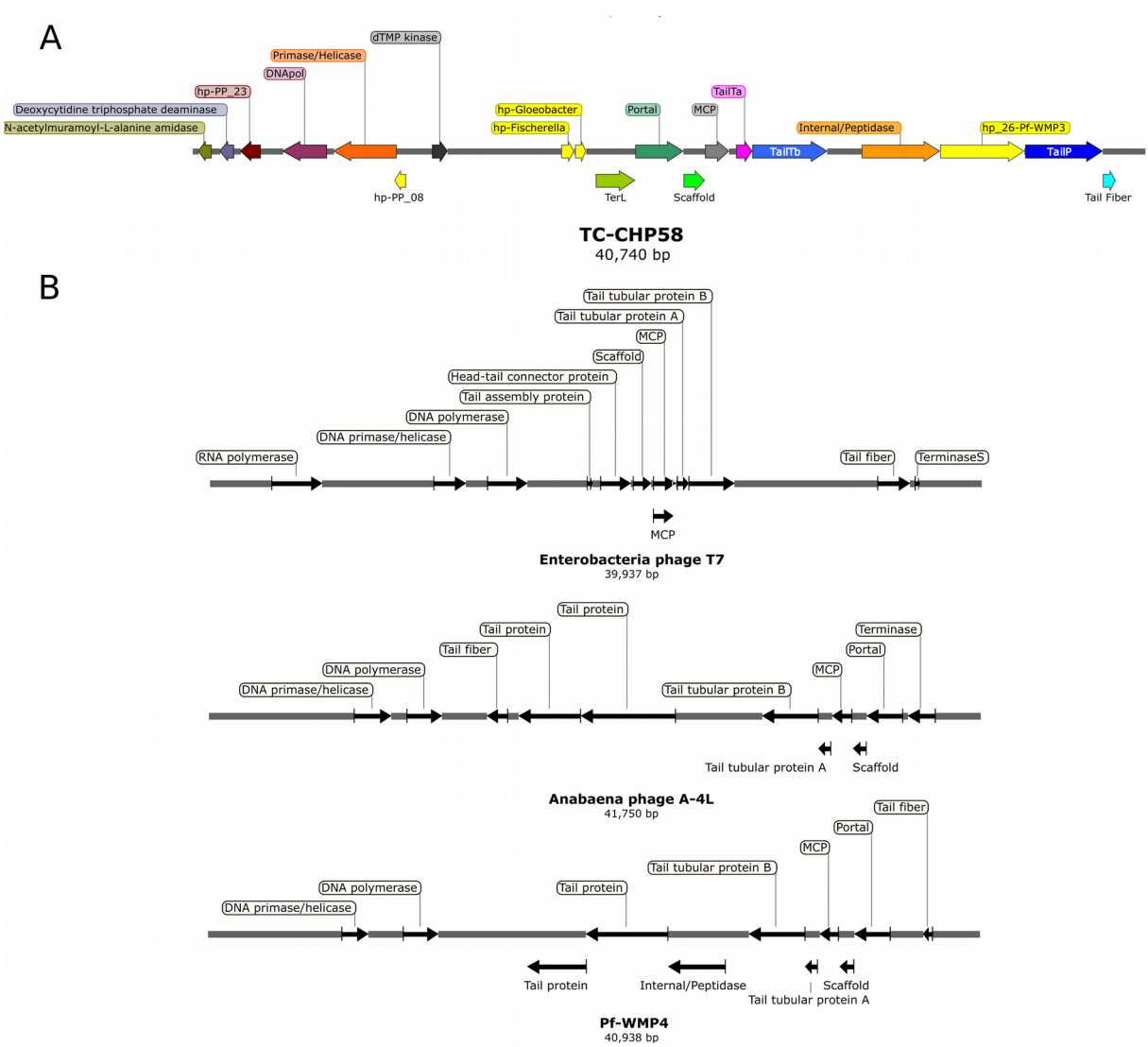


Figure 4. A) Genomic organization of the Thermophilic Cyanophage CHP58 (TC-CHP58). Arrows indicate the size, position, and orientation of annotated ORFs, with predicted functions or homologues (e.g. DNApol, DNA polymerase; TailY a/b, tail tubular protein a/b; MCP, major capsid protein; TerL, large terminase subunit; TailP, Tail protein; hp-PP_08/23, homologous to hypothetical proteins 08/23 from cyanophage PP; hp-*Fischerella/Gloeobacter*, homologous to hypothetical proteins in *Fischerella* sp. PCC 9605/*Gloeobacter kilaueensis*). B) Genomic organization of Enterobacteria phage T7, Anabaena phage A4L, and Pf-WMP4. Arrows indicate the size, position, and orientation of viral core ORFs.

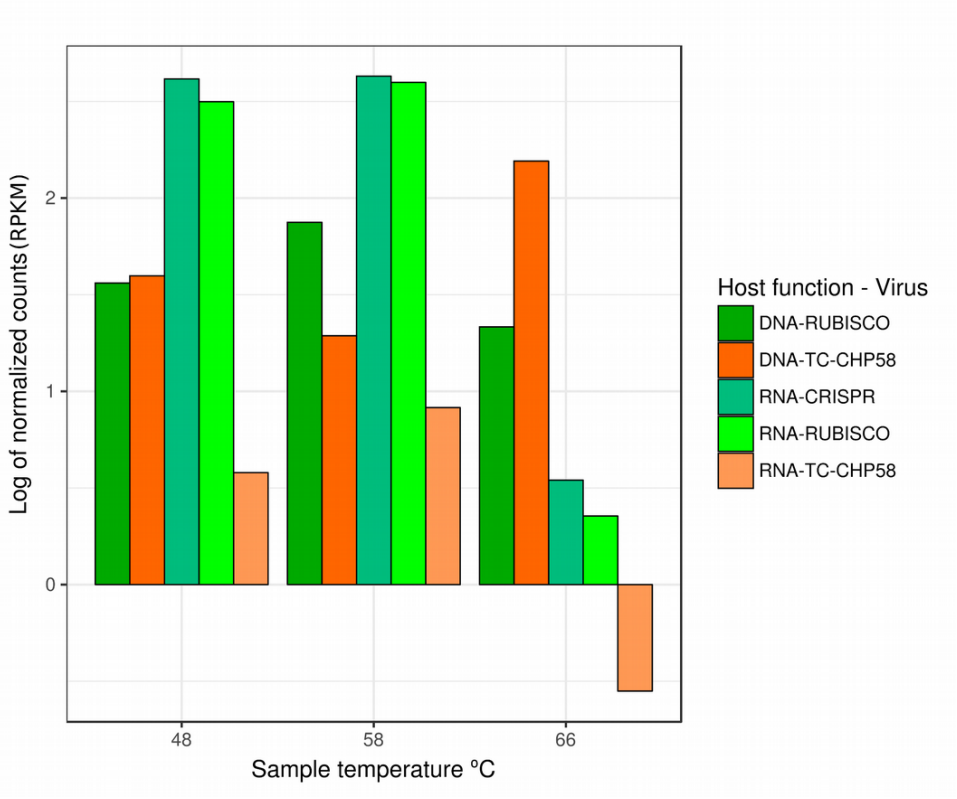


Figure 5. Relative abundance and transcriptomic expression of *Mastigocladus* spp. CRISPR systems and TC-CHP58. Abundance and expression of *Mastigocladus* RUBISCO was used as reference of the cyanobacterial presence and metabolic activity. Only specific CRISPR *loci* with proto-spacers in TC-CHP58 were fully quantified for each temperature. For improved visualization, counts are represented as Log of reads per kilobase million (RPKM).

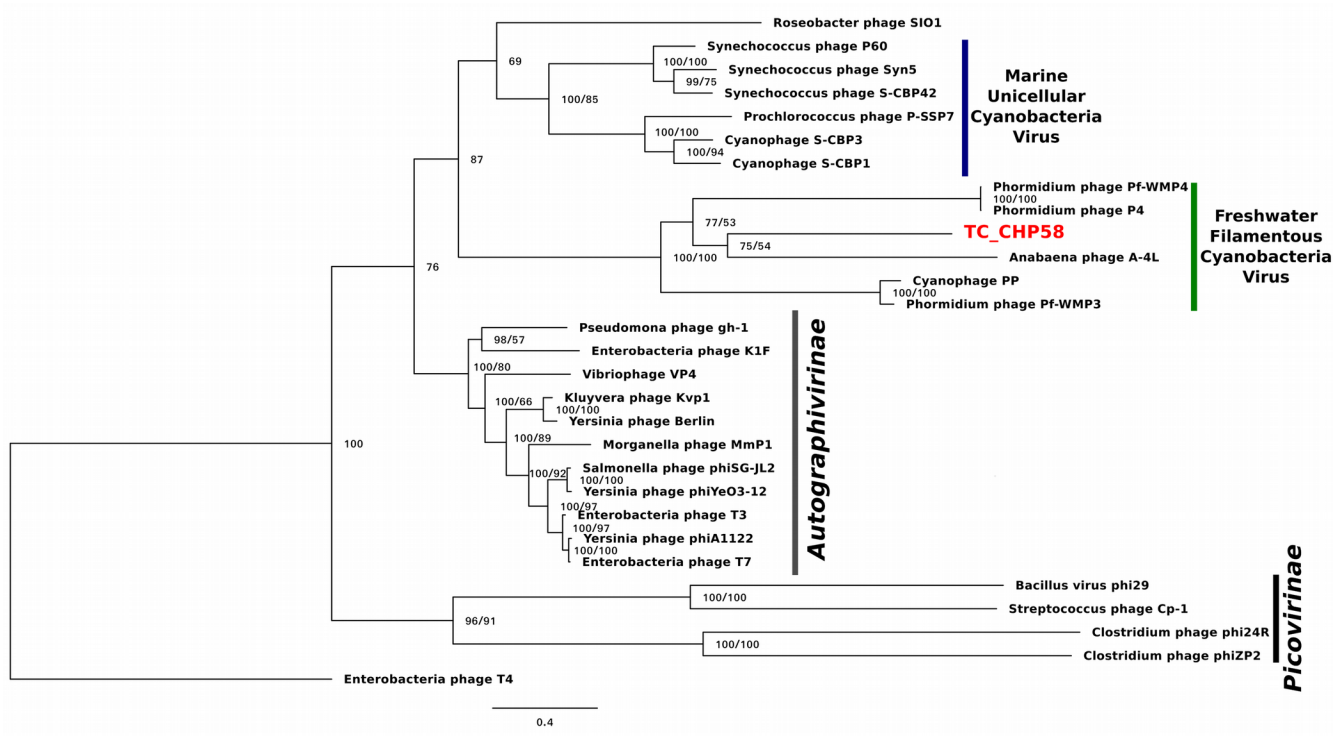


Figure 6. Bayesian inference phylogenetic reconstruction of DNA polymerase I protein of TC-CHP58. Numbers indicate Bayesian posterior probabilities as percentage/ultra-fast bootstrap values. Only UFBoot values over 80 and Bayesian PP over 50 are shown. The sequence characterized in the present study is reported in bold letters. Scale bar: 0.4 amino acid substitutions per site.

Tables

Table1. Blastp analysis of predicted CDS from TC-CHP58 of known function against NCBI RefSeq (Release 75) and NR databases.

Query sequence ID	Subject sequence ID	Identity %	E-value	Bit Score
TC-CHP58_sequence	KF598865.1| [Cyanophage PP]	93%	0.2	54.7
TC-CHP58_CDS1	YP_009042789.1| DNA polymerase [Anabaena phage A-4L]	29.03	4.00E-60	223
TC-CHP58_CDS3	YP_009042786.1| DNA primase/helicase [Anabaena phage A-4L]	25.21	2.00E-28	131
TC-CHP58_CDS4	YP_008766966.1| hypothetical protein PP_08 [Cyanophage PP]	29.7	0.0006	47
TC-CHP58_CDS7	WP_026824764.1| dTMP kinase [Exiguobacterium marinum]	30.61	4.00E-22	99.4
TC-CHP58_CDS13	WP_026731322.1| hypothetical protein [Fischerella sp. PCC 9605]	34.55	7.00E-12	69.3
TC-CHP58_CDS14	WP_023172199.1| hypothetical protein [Gloeobacter kilaueensis]	42.31	4E-05	48.1
TC-CHP58_CDS15	YP_008766995.1| terminase [Cyanophage PP]	44.39	2.00E-150	456
TC-CHP58_CDS16	YP_001285799.1| portal protein [Phormidium phage Pf-WMP3]	42.96	0	551
TC-CHP58_CDS17	YP_009042804.1| scaffold protein [Anabaena phage A-4L]	30.69	1E-07	60.8
TC-CHP58_CDS18	YP_008766991.1| capsid protein [Cyanophage PP]	48.14	4.00E-109	335
TC-CHP58_CDS20	YP_009042802.1| tail tubular protein A [Anabaena phage A-4L]	29.52	3.00E-28	116
TC-CHP58_CDS21	YP_001285795.1| tail tubular protein B [Phormidium phage Pf-WMP3]	36.49	0	630
TC-CHP58_CDS24	YP_009042798.1| internal protein [Anabaena phage A-4L]	29.86	5.00E-41	174
TC-CHP58_CDS25	YP_001285791.1| PfWMP3_26 [Phormidium phage Pf-WMP3]	28.26	5.00E-23	117
TC-CHP58_CDS26	YP_009042796.1| tail protein [Anabaena phage A-4L]	24.8	4.00E-89	322
TC-CHP58_CDS32	WP_038085449.1| N-acetylmuramoyl-L-alanine amidase [Tolypothrix bouteillei]	46.29	3.00E-46	160
TC-CHP58_CDS35	WP_043587103.1| deoxycytidine triphosphate deaminase [Diplosphaera colitermitum]	44.9	2.00E-48	167
TC-CHP58_CDS37	YP_008766981.1| hypothetical protein PP_23 [Cyanophage PP]	29.92	1.00E-21	101

Table 2. CRISPR *loci* at each temperature detailed information and SNVs analysis for TC-CHP58 proto-spacer including alleles frequency and SNV coding effect.

T° Sample	Virotope Sequence	Viral target	CRISPR *loci*	Proto-spacer Start	Proto-spacer End	Mismatch position	SNV position	Alleles *	Frequency of in CRISPR allele	SNV effect	Codon change
48	ACCTTTCAGACCTAACTCTA AAGTTACTATCACAGAT	internal protein-M23-peptidase	48_CRISPR_2_NODE_1554	26558	26594	26564; 26567; 26582	--	--	--	--	--
58	AGAAGTTTTTCTTCGCCAAG ATATATGGTGCTGGTCTAA	DNA polymerase	58_CRISPR_10_NODE_13413	282	320	--	282; 292; 313	G/T; C/A; G/A	0.079; 0.552; 0.549	All silent	CGA/AGA; GGG/CGT; TTC/TTT
58	GTGTTGGTGCTCTTGGAGTA CCGTTCAGAATAGGT	Hypothetical protein	58_CRISPR_10_NODE_13413	35908	35942	--	35908	G/A	0.052	Silent	GGC/GGT
58	AGTTGTGCCCCTTGAGCTAG AGAATTTGCTGCACCT	internal protein-M23-peptidase	58_CRISPR_10_NODE_13413	24692	24727	--	--	--	--	--	--
58	TAAACTGGTCGGGATTGTGT ACATTCCATGCACTC	NC	58_CRISPR_10_NODE_13413	8740	8774	--	8753	C/G	0.53	--	--
58	ACTATCTGATCAAACCGGGG CTACACGGTAAATCGTTAGA	Tail fiber protein	58_CRISPR_10_NODE_13413	36649	36688	36650	36675	C/T	0.522	A/V	GCT/GTT
58	ACCTTTCAGACCTAACTCTA AAGTTACTATCACAGAT	internal protein-M23-peptidase	58_CRISPR_5_NODE_1091	26558	26594	26565; 26567; 26582	--	--	--	--	--
58	CCCAACAACGTCTAAATAAA TCTTTCTATGATATGC	Hypothetical protein	58_CRISPR_8_NODE_4438	24219	24254	24226; 24229; 24238	--	--	--	--	--
58	AATACGGTTGTAGTACTCTTG AAGAGGTGTTACCG	Hypothetical protein	58_CRISPR_8_NODE_4438	30971	31005	30972	30978	G/A	0.588	Silent	ACG/ACA
58	GAAAGGGTAAGGTGTCAAA ATTGGGATTATTAGTGTTAG	internal protein-M23-peptidase	58_CRISPR_8_NODE_4438	27172	27210	--	27180; 27181; 27186	T/A; C/A; A/C	0.626; 0.613; 0.575	S/T; S/Y; T/P	TCT/ACT; TCT/TAT; ACC/CCC
58	GCATTAATCGCGGGGTTAGG GTGATACCACCTA	tail protein	58_CRISPR_8_NODE_4438	34211	34243	34214; 34226; 34241	--	--	--	--	--
58	TAGCTTAACATTACCACAGG GGATAAGCTGTTGTATATCC	deoxycytidine triphosphate deaminase	58_CRISPR_9_NODE_4711	38225	38264	38264	38225; 38261	C/T G/A	0.057; 0.046	All silent	CTG/CTA; GAC;GAT
58	GACTTGATCTTTTCCGCTTTC TTGTAGCGCAGTATCTT	DNA polymerase	58_CRISPR_9_NODE_4711	668	705	--	671; 673	C/T; T/G	0.031; 0.040	R/K; Silent	AGG/AAG
58	ACGGGGTTGATCTTCCCCGC GAAGTGGTTGTCACCGAAT	dTMP kinase	58_CRISPR_9_NODE_4711	6338	6376	6375	6350	G/C	0.575	K/N	AAG/AAC
58	AAATACATCCCCCACTTTAG GAGGTAACCCCAC	Hypothetical protein	58_CRISPR_9_NODE_4711	36127	36159	--	--	--	--	--	--
58	ACAGCGAAAGCAATTTGTCT CTGAGGCTAACAAGTT	internal protein-M23-peptidase	58_CRISPR_9_NODE_4711	25742	25777	25776	--	--	--	--	--
58	GTCGTATCTCAATGTACTCTT TGTAGTCTTTCCA	internal protein-M23-peptidase	58_CRISPR_9_NODE_4711	25917	25950	--	25946	C/A	0.041	Silent	ATC/ATA
58	CAATCACACCTAACCCCATA GGTGACCGCACAACA	portal protein	58_CRISPR_9_NODE_4711	15920	15954	--	15941; 15953	A/G; A/T	0.651; 0.055	All silent	GGA/GGG; ATA/ATT
58	TAGCTGATTGGAAAGCAGAC GCTGGATTATTACAC	tail protein	58_CRISPR_9_NODE_4711	33859	33893	--	--	--	--	--	--
66	ATCTGTGATAGTAACTTTAGA GTTAGGTCTGAAAGGT	internal protein-M23-peptidase	66_CRISPR_2_NODE_1045	26558	26594	26566; 26567; 26582	--	--	--	--	--
66	TTAGACCAGCACCATATATCT TGGCGAAGAAAAACTTCT	DNA polymerase	66_CRISPR_3_NODE_1491	282	320	282	292; 313	C/A; G/A	0.437; 0.024	All silent	GGG/CGT; TTC/TTT
66	ACCTATTCTGAACGGTACTC CAAGAGCACCAACAC	Hypothetical protein	66_CRISPR_3_NODE_1491	35908	35942	35908	--	--	--	--	--
66	AGGTGCAGCAAATTCTCTAG CTCAAGGGGCACAACT	internal protein-M23-peptidase	66_CRISPR_3_NODE_1491	24692	24727	--	--	--	--	--	--
66	GAGTGCATGGAATGTACACA ATCCCGACCAGTTTA	NC	66_CRISPR_3_NODE_1491	8740	8774	--	8753	C/G	0.051	--	--
66	TCTAACGATTTACCGTGTAGC CCCGGTTTGATCAGATAGT	Tail fiber protein	66_CRISPR_3_NODE_1491	36649	36688	36650	36675	C/T	0.056	A/V	GCT/GTT

Supplementary Material

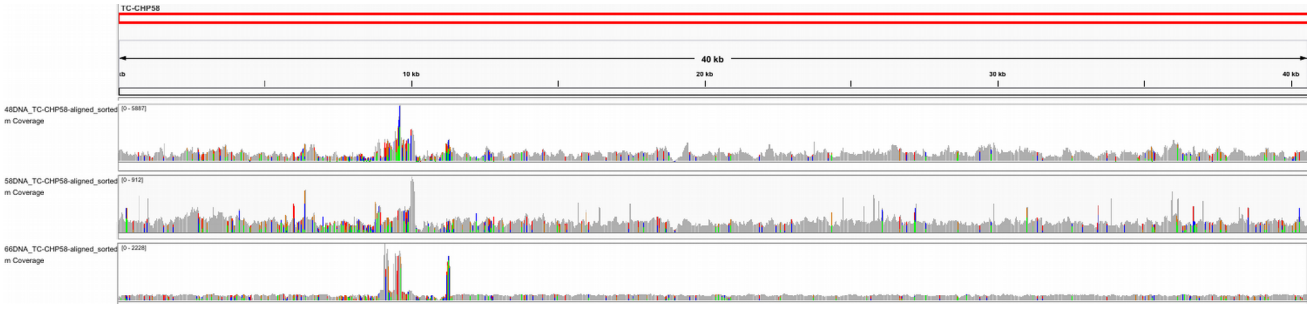
Active crossfire between Cyanobacteria and Cyanophages in phototrophic mat communities within hot springs

Sergio Guajardo-Leiva, Carlos Pedrós-Alió, Oscar Salgado, Fabián Pinto, and Beatriz Díez

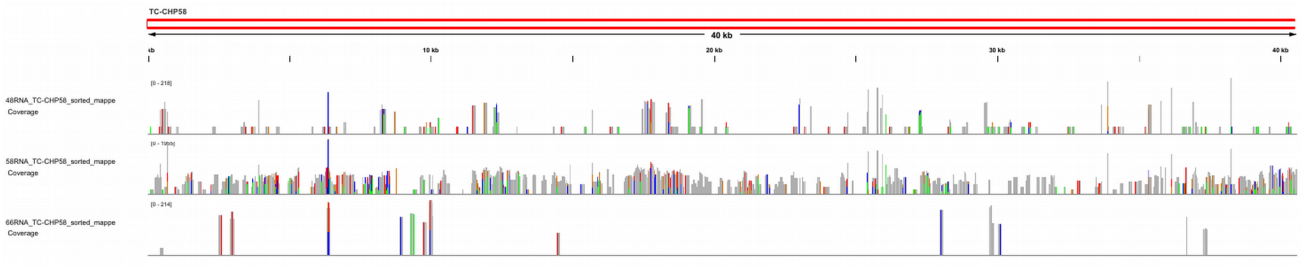
*** Correspondence:** Beatriz Díez: bdiez@bio.puc.cl

Supplementary Figures and Tables

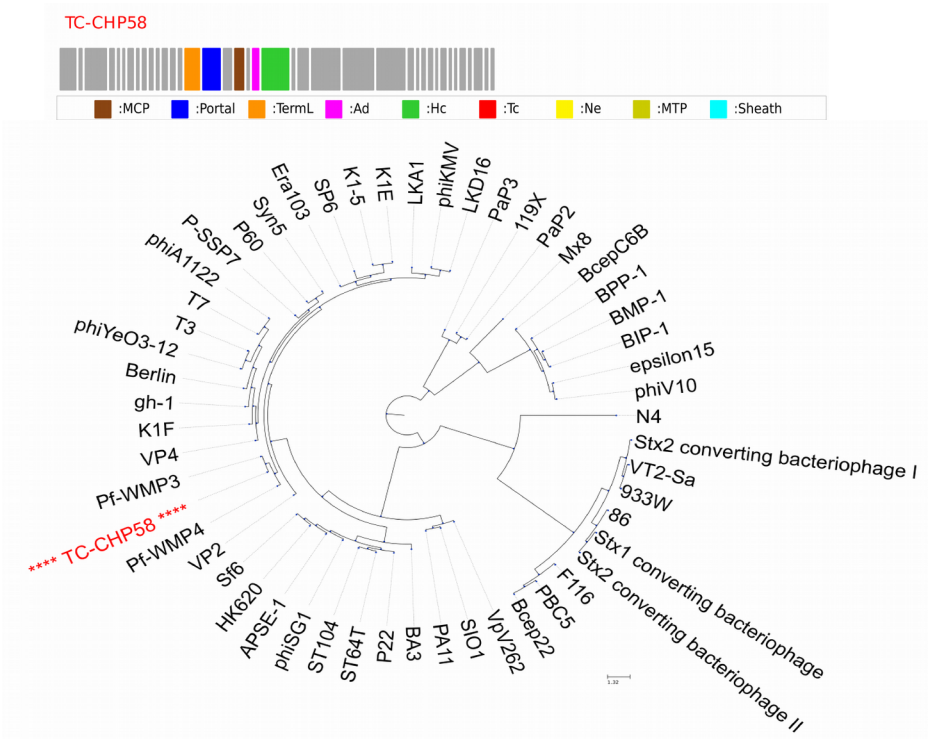
Supplementary Figures



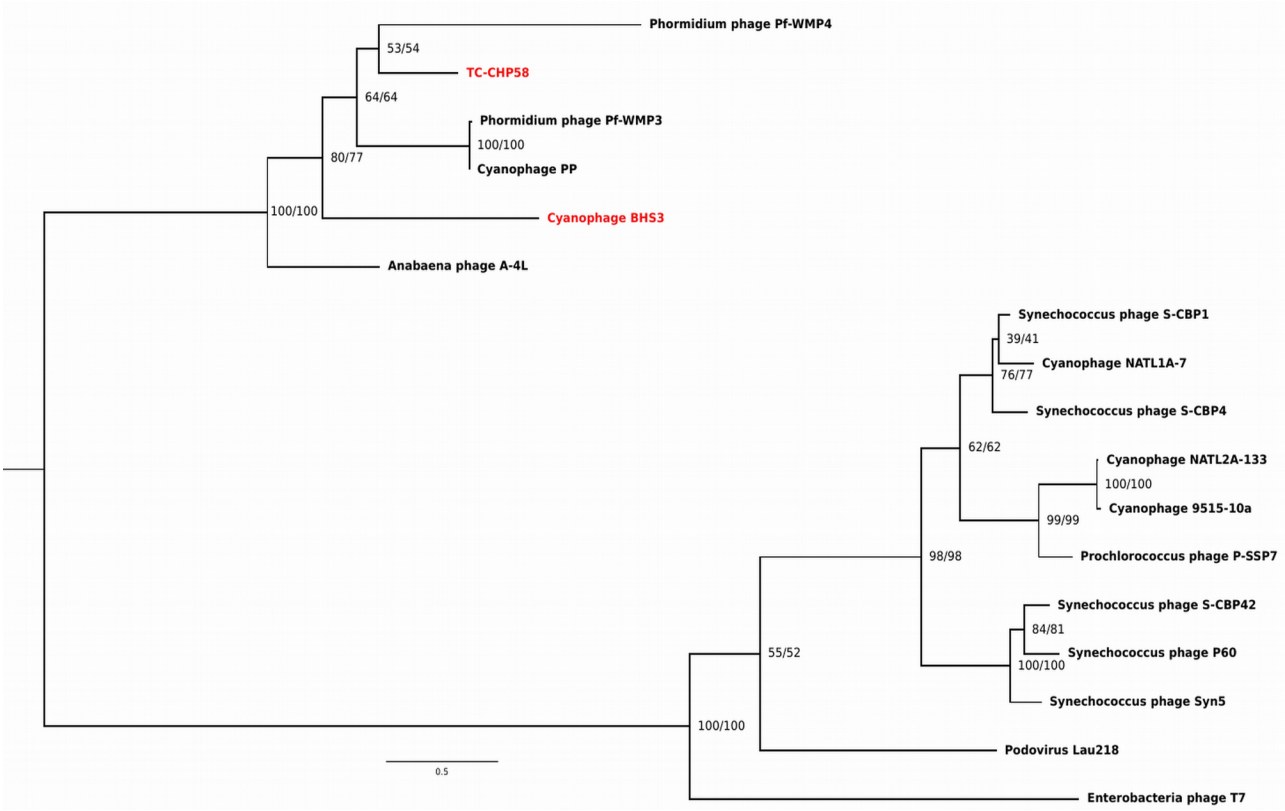
Supplementary Figure S1. Reads recruitment to TC-CHP58 genome from metagenomes of Porcelana hot spring temperature gradient. Grey lines represent zones of perfect match and colored lines represent different ratios of nucleotide mismatched positions.



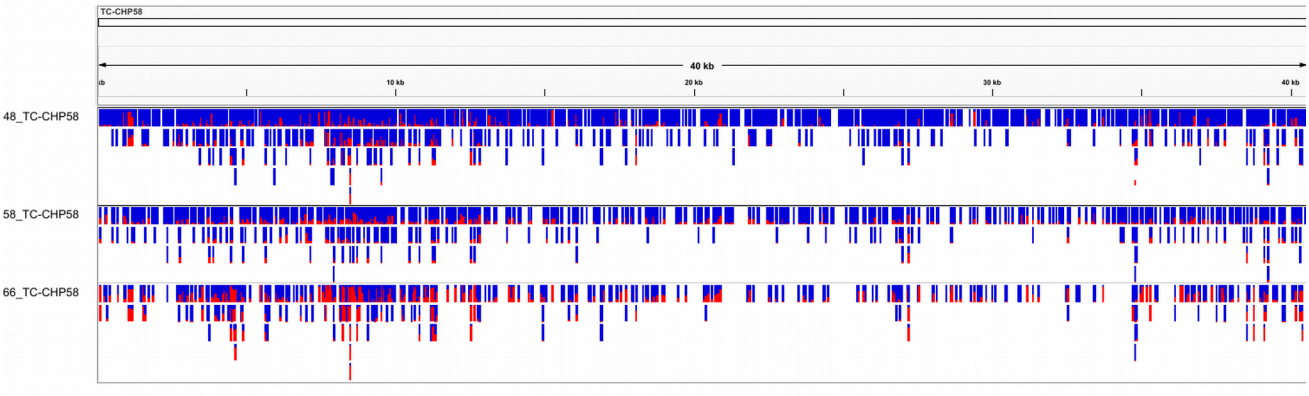
Supplementary Figure S2. Reads recruitment to TC-CHP58 genome from metatrancriptomes of Porcelana hot spring temperature gradient. Grey lines represent zones of perfect match and colored lines represent different ratios of nucleotide mismatched positions. White zones represent absence of mapping.



Supplementary Figure S3. TC-CHP58 VIRFAM neck protein organization analyses. Upper panel show a schematic representation of TC-CHP58 gene organization. Bottom panel show a hierarchical clustering of neck proteins of TC-CHP58 and other related viruses.



Supplementary Figure S4. Maximum likelihood phylogenetic gene tree of Major Capsid protein. Numbers indicate Ultra fast bootstrap (UFBoot) values. Only UFBoot values over 50 are show. The sequence characterized in the present work is reported in red bold letters. Scale bar: 0.6 amino acid substitutions per site.



Supplementary Figure S5. SNVs calling along TC-CHP58 genome for TC-CHP58 populations over Porcelana temperature gradient. Red and blue lines represent the ratio of different alleles for each SNV.

Supplementary Tables

Supplementary Table S1. Summary information about sequencing depth, quality filtering, read mapping and assembly of Porcelana metagenomes an metatranscriptomes.

Sample	Raw reads (10E6)	Raw bases (10E6)	Reads (10E6) after quality filter	Bases (10E6) after quality filter	Assembled reads (10E6)	Assembled bases (10E6)	VNA aligned reads (10E6)	VNA aligned bases (10E6)	Bacteria 16S aligned reads	Archaea 16S aligned reads	Eukarya 18s aligned reads
48DNA	336.8	42448.92	279.37	21886.89	46.43	5585.51	1.1	123.23	147241	1981	683
58DNA	140.12	17655.26	118.83	10263.5	34.83	4176.35	0.559	62.36	90911	95	473
66DNA	10.5	1130.51	7.143	684.32	6.28	604.54	0.044	5.39	6928	4	396
48RNA	15.5	1953.23	8.4	1035.14	--	--	0.312	38.9	--	--	--
58RNA	38.5	4850.86	22.71	2819.89	--	--	0.472	59.05	--	--	--
66RNA	149.2	18805.33	52.12	6866.81	--	--	0.185	23.3	--	--	--

Supplementary Table S2. Accession numbers of sequences used in phylogenetic analyses of DNA Polymerase, and Major Capsid of TC-CHP58

DNApol Accession number	Organism	Capsid Accession number	Organism
YP_001285436.1	Synechococcus phage Syn5	NP_041998.1	Enterobacteria phage T7
AAG02598.1	Roseobacter phage SIO1	YP_001285448.1	Synechococcus phage Syn5
AAL73268.1	Synechococcus phage P60	YP_001285797.1	Phormidium phage Pf-WMP3
AAO73157.1	Pseudomonad phage gh-1	YP_005087431.1	Cyanophage 9515-10a
ABG75891.1	Phormidium phage P4	YP_005087453.1	Cyanophage NATL1A-7
ABU50233.1	Cyanophage S-CBP1	YP_005087541.1	Cyanophage NATL2A-133
ABU50234.1	Cyanophage S-CBP3	YP_008766991.1	Cyanophage PP
ABU50235.1	Synechococcus phage S-CBP42	YP_009042168.1	Podovirus Lau218
CAA24412.1	Enterobacteria phage T7	YP_009042803.1	Anabaena phage A-4L
CAB63614.1	Yersinia phage phiYeO3-12	YP_009103190.1	Synechococcus phage S-CBP1
CAC86283.1	Enterobacteria phage T3	YP_009103797.1	Synechococcus phage S-CBP4
NP_044817.1	Streptococcus phage Cp-1	YP_009173806.1	Synechococcus phage P60
NP_049662.1	Enterobacteria phage T4	YP_009220191.1	Synechococcus phage S-CBP42
NP_848283.1	Yersinia phage phiA1122	YP_214206.1	Prochlorococcus phage P-SSP7
YP_001285777.1	Phormidium phage Pf-WMP3	YP_762667.1	Phormidium phage Pf-WMP4
YP_001949769.1	Salmonella phage phiSG-JL2	ASV43892.1	Cyanophage BHS3
YP_002004529.1	Bacillus virus phi29	--	--
YP_002048647.1	Morganella phage MmP1	--	--
YP_002308401.1	Kluyvera phage Kvp1	--	--
YP_006355438.1	Prochlorococcus phage P-SSP7	--	--
YP_006488652.1	Clostridium phage phiZP2	--	--
YP_007006982.1	Clostridium phage phi24R	--	--
YP_008766972.1	Cyanophage PP	--	--
YP_009042789.1	Anabaena phage A-4L	--	--
YP_249581.1	Vibriophage VP4	--	--
YP_338108.1	Enterobacteria phage K1F	--	--
YP_762649.1	Phormidium phage Pf-WMP4	--	--
YP_919001.1	Yersinia phage Berlin	--	--

Supplementary Table S3. Relative abundances (%) of Caudovirales community at the host Phylum level, obtained from shotgun sequences in metagenomes (DNA) and metatranscriptomes (RNA), standardized by the total number of reads from each temperature sample.

Temperature ºC	Family	Host-Virus	Percent	Temperature ºC	Family	Host-Virus	Percent	Temperature ºC	Family	Host-Virus	Percent
DNA-48 ºC	Siphoviridae	Actinobacteria	40.2	DNA-48 ºC	Podoviridae	Cyanophyceae	30.5	DNA-48 ºC	Myoviridae	Gammaproteobacteria	36.1
DNA-48 ºC	Siphoviridae	Bacilli	18.7	DNA-48 ºC	Podoviridae	Flavobacteriia	21.2	DNA-48 ºC	Myoviridae	Cyanophyceae	30
DNA-48 ºC	Siphoviridae	Gammaproteobacteria	36.2	DNA-48 ºC	Podoviridae	Gammaproteobacteria	34.7	DNA-48 ºC	Myoviridae	Bacilli	22
DNA-48 ºC	Siphoviridae	Flavobacteriia	2	DNA-48 ºC	Podoviridae	Bacilli	4.1	DNA-48 ºC	Myoviridae	Actinobacteria	3.5
DNA-48 ºC	Siphoviridae	Alphaproteobacteria	0.6	DNA-48 ºC	Podoviridae	Actinobacteria	3.5	DNA-48 ºC	Myoviridae	Clostridia	2.5
DNA-48 ºC	Siphoviridae	Verrucomicrobia	0.8	DNA-48 ºC	Podoviridae	BetaProteobacteria	3.5	DNA-48 ºC	Myoviridae	Betaproteobacteria	3.1
DNA-48 ºC	Siphoviridae	Methanobacteria	0.2	DNA-48 ºC	Podoviridae	Enviromental	2.4	DNA-48 ºC	Myoviridae	Alphaproteobacteria	2.7
DNA-48 ºC	Siphoviridae	Clostridia	0.1	DNA-58 ºC	Podoviridae	Cyanophyceae	50.5	DNA-58 ºC	Myoviridae	Gammaproteobacteria	24.8
DNA-48 ºC	Siphoviridae	BetaProteobacteria	1.2	DNA-58 ºC	Podoviridae	Flavobacteriia	22.8	DNA-58 ºC	Myoviridae	Cyanophyceae	44.8
DNA-58 ºC	Siphoviridae	Actinobacteria	52.2	DNA-58 ºC	Podoviridae	Gammaproteobacteria	18.8	DNA-58 ºC	Myoviridae	Bacilli	19.7
DNA-58 ºC	Siphoviridae	Bacilli	25.6	DNA-58 ºC	Podoviridae	Bacilli	3.6	DNA-58 ºC	Myoviridae	Actinobacteria	3.1
DNA-58 ºC	Siphoviridae	Gammaproteobacteria	17.5	DNA-58 ºC	Podoviridae	Actinobacteria	2.1	DNA-58 ºC	Myoviridae	Clostridia	2.6
DNA-58 ºC	Siphoviridae	Flavobacteriia	3	DNA-58 ºC	Podoviridae	BetaProteobacteria	1.3	DNA-58 ºC	Myoviridae	Betaproteobacteria	2.7
DNA-58 ºC	Siphoviridae	Alphaproteobacteria	0.8	DNA-58 ºC	Podoviridae	Enviromental	0.9	DNA-58 ºC	Myoviridae	Alphaproteobacteria	2.4
DNA-58 ºC	Siphoviridae	Verrucomicrobia	0.4	DNA-66 ºC	Podoviridae	Cyanophyceae	23.2	DNA-66 ºC	Myoviridae	Gammaproteobacteria	33.8
DNA-58 ºC	Siphoviridae	Methanobacteria	0.2	DNA-66 ºC	Podoviridae	Flavobacteriia	25.5	DNA-66 ºC	Myoviridae	Cyanophyceae	27.9
DNA-58 ºC	Siphoviridae	Clostridia	0.2	DNA-66 ºC	Podoviridae	Gammaproteobacteria	38.9	DNA-66 ºC	Myoviridae	Bacilli	24.6
DNA-58 ºC	Siphoviridae	BetaProteobacteria	0.1	DNA-66 ºC	Podoviridae	Bacilli	5.7	DNA-66 ºC	Myoviridae	Actinobacteria	4
DNA-66 ºC	Siphoviridae	Actinobacteria	52.6	DNA-66 ºC	Podoviridae	Actinobacteria	2.4	DNA-66 ºC	Myoviridae	Clostridia	4
DNA-66 ºC	Siphoviridae	Bacilli	25.7	DNA-66 ºC	Podoviridae	BetaProteobacteria	3.1	DNA-66 ºC	Myoviridae	Betaproteobacteria	3.1
DNA-66 ºC	Siphoviridae	Gammaproteobacteria	16.8	DNA-66 ºC	Podoviridae	Enviromental	1.3	DNA-66 ºC	Myoviridae	Alphaproteobacteria	2.6
DNA-66 ºC	Siphoviridae	Flavobacteriia	3.2	RNA-48 ºC	Podoviridae	Cyanophyceae	95	RNA-48 ºC	Myoviridae	Gammaproteobacteria	1.5
DNA-66 ºC	Siphoviridae	Alphaproteobacteria	0.8	RNA-48 ºC	Podoviridae	Flavobacteriia	4.3	RNA-48 ºC	Myoviridae	Cyanophyceae	96
DNA-66 ºC	Siphoviridae	Verrucomicrobia	0.3	RNA-48 ºC	Podoviridae	Gammaproteobacteria	0.3	RNA-48 ºC	Myoviridae	Bacilli	1.7
DNA-66 ºC	Siphoviridae	Methanobacteria	0.2	RNA-48 ºC	Podoviridae	Bacilli	0	RNA-48 ºC	Myoviridae	Actinobacteria	0.2
DNA-66 ºC	Siphoviridae	Clostridia	0.2	RNA-48 ºC	Podoviridae	Actinobacteria	0.1	RNA-48 ºC	Myoviridae	Clostridia	0.4
DNA-66 ºC	Siphoviridae	BetaProteobacteria	0.2	RNA-48 ºC	Podoviridae	BetaProteobacteria	0.2	RNA-48 ºC	Myoviridae	Betaproteobacteria	0.1
RNA-48 ºC	Siphoviridae	Actinobacteria	66.5	RNA-48 ºC	Podoviridae	Enviromental	0	RNA-48 ºC	Myoviridae	Alphaproteobacteria	0.2
RNA-48 ºC	Siphoviridae	Bacilli	24.6	RNA-58 ºC	Podoviridae	Cyanophyceae	84	RNA-58 ºC	Myoviridae	Gammaproteobacteria	5
RNA-48 ºC	Siphoviridae	Gammaproteobacteria	7.7	RNA-58 ºC	Podoviridae	Flavobacteriia	14.4	RNA-58 ºC	Myoviridae	Cyanophyceae	88.9
RNA-48 ºC	Siphoviridae	Flavobacteriia	0.7	RNA-58 ºC	Podoviridae	Gammaproteobacteria	0.7	RNA-58 ºC	Myoviridae	Bacilli	4
RNA-48 ºC	Siphoviridae	Alphaproteobacteria	0.5	RNA-58 ºC	Podoviridae	Bacilli	0.3	RNA-58 ºC	Myoviridae	Actinobacteria	0.5
RNA-48 ºC	Siphoviridae	Verrucomicrobia	0	RNA-58 ºC	Podoviridae	Actinobacteria	0.2	RNA-58 ºC	Myoviridae	Clostridia	1
RNA-48 ºC	Siphoviridae	Methanobacteria	0	RNA-58 ºC	Podoviridae	BetaProteobacteria	0.3	RNA-58 ºC	Myoviridae	Betaproteobacteria	0.2
RNA-48 ºC	Siphoviridae	Clostridia	0	RNA-58 ºC	Podoviridae	Enviromental	0.1	RNA-58 ºC	Myoviridae	Alphaproteobacteria	0.3
RNA-48 ºC	Siphoviridae	BetaProteobacteria	0	RNA-66 ºC	Podoviridae	Cyanophyceae	14.6	RNA-66 ºC	Myoviridae	Gammaproteobacteria	33.9
RNA-58 ºC	Siphoviridae	Actinobacteria	69.7	RNA-66 ºC	Podoviridae	Flavobacteriia	43.9	RNA-66 ºC	Myoviridae	Cyanophyceae	19.6
RNA-58 ºC	Siphoviridae	Bacilli	21.1	RNA-66 ºC	Podoviridae	Gammaproteobacteria	20.4	RNA-66 ºC	Myoviridae	Bacilli	27
RNA-58 ºC	Siphoviridae	Gammaproteobacteria	8	RNA-66 ºC	Podoviridae	Bacilli	12.2	RNA-66 ºC	Myoviridae	Actinobacteria	5.5
RNA-58 ºC	Siphoviridae	Flavobacteriia	0.6	RNA-66 ºC	Podoviridae	Actinobacteria	4.2	RNA-66 ºC	Myoviridae	Clostridia	7.9
RNA-58 ºC	Siphoviridae	Alphaproteobacteria	0.3	RNA-66 ºC	Podoviridae	BetaProteobacteria	4.2	RNA-66 ºC	Myoviridae	Betaproteobacteria	2.5
RNA-58 ºC	Siphoviridae	Verrucomicrobia	0.1	RNA-66 ºC	Podoviridae	Enviromental	0.6	RNA-66 ºC	Myoviridae	Alphaproteobacteria	3.5
RNA-58 ºC	Siphoviridae	Methanobacteria	0.2	--	--	--	--	--	--	--	--
RNA-58 ºC	Siphoviridae	Clostridia	0	--	--	--	--	--	--	--	--
RNA-58 ºC	Siphoviridae	BetaProteobacteria	0	--	--	--	--	--	--	--	--

RNA-66 ºC	Siphoviridae	Actinobacteria	43.9	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Bacilli	28.3	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Gammaproteobacteria	24.1	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Flavobacteriia	1.3	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Alphaproteobacteria	0.4	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Verrucomicrobia	0	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Methanobacteria	0.2	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	Clostridia	1.7	--	--	--	--	--	--	--	--
RNA-66 ºC	Siphoviridae	BetaProteobacteria	0.1	--	--	--	--	--	--	--	--

Supplementary Table S4. Relative abundance and transcriptomic expression (RPKM) of *Mastigocladus* sp. CRISPR *loci* and TC-CHP58. Reads were normalized by size and depth of sequencing. Abundance and expression of *Mastigocladus* RUBISCO was used as reference of the cyanobacterial presence and metabolic activity.

Sample T ºC	RUBISCO-DNA-RPKM	DNA_TC-CP58-RPKM	RUBISCO-RNA-RPKM	CRISPR-RNA-RPKM	RNA_TC-CHP58-RPKM
48	36.31	39.58	315.46	413.34	3.79
58	74.86	19.40	397.37	427.61	8.24
66	21.55	155.47	2.26	3.47	0.28

Supplementary Table S5. Number of Single nucleotide variants in TC-CHP58 ORFs at different metagenomes of Porcelana temperature gradient.

`ORFs	Start	End	48_SNVs	58_SNVs	66_SNVs	Length
ORF_1-DNA polymerase	40528	1638	36	23	17	1850
ORF_2-Hypothetical protein	1619	1909	11	4	3	290
ORF_3-DNA primase/helicase	1960	4470	129	79	50	2510
ORF_4-Hypothetical protein	4553	5041	39	15	27	488
ORF_5-Hypothetical protein	5038	5220	12	2	2	182
ORF_6-Hypothetical protein	5844	6023	12	4	2	179
ORF_7-dTMP kinase	6246	6839	36	22	14	593
ORF_8-Hypothetical protein	6960	7211	12	12	4	251
ORF_9-Hypothetical protein	7589	7987	48	37	29	398
ORF_10-Hypothetical protein	8158	8556	61	35	38	398
ORF_11-Hypothetical protein	11435	11704	7	1	0	269
ORF_12-Hypothetical protein	11722	12273	18	16	4	551
ORF_13-Hypothetical protein	12292	12780	19	20	11	488
ORF_14-Hypothetical protein	12820	13173	12	5	2	353
ORF_15-Terminase	13177	14886	58	18	13	1709
ORF_16-Portal protein	14886	16931	65	41	34	2045
ORF_17-Scaffold protein	16928	17842	26	9	8	914
ORF_18-Major capsid protein	17862	18896	28	18	14	1034
ORF_19-Tail tubular protein A	18977	19198	8	1	1	221
ORF_20-Tail tubular protein B	19195	19887	19	5	2	692
ORF_21-Hypothetical protein	19889	23071	85	38	21	3182
ORF_22-Hypothetical protein	23088	23357	4	2	0	269
ORF_23-Hypothetical protein	23360	24559	33	17	10	1199
ORF_24-Internal protein/peptidase	24552	27917	104	62	23	3365
ORF_25-Hypothetical protein	27919	31536	78	35	25	3617

ORF_26-Tail protein	31542	34883	94	58	25	3341
ORF_27-Hypothetical protein	34909	35472	22	15	7	563
ORF_28-Hypothetical protein	35605	35790	3	3	0	185
ORF_29-Hypothetical protein	35848	36204	13	7	2	356
ORF_30-Tail Fiber	36245	36679	25	14	7	434
ORF_31-Hypothetical protein	36642	36806	6	5	3	164
ORF_32-N-acetylmuramoyl-L-alanine amidase	36884	37444	19	12	6	560
ORF_33-Hypothetical protein	37428	37634	11	8	5	206
ORF_34-Hypothetical protein	37637	37828	13	9	7	191
ORF_35-Deoxycytidine triphosphate deaminase	37835	38419	13	8	3	584
ORF_36-Hypothetical protein	38434	38697	11	8	7	263
ORF_37-Hypothetical protein	38694	39542	36	19	20	848
ORF_38-Hypothetical protein	39611	39880	9	11	7	269
ORF_39-Hypothetical protein	39892	40182	12	8	7	290

1288

1289

CHAPTER 2

Killing the winner and piggybacking the cheater: lytic and lysogenic viral communities in hot springs phototrophic mats.

Killing the winner and piggybacking the cheater: lytic and lysogenic viral communities in hot springs phototrophic mats.

Sergio Guajardo-Leiva[1], Oscar Salgado[1], and Beatriz Díez[1,2]

[1]Department of Molecular Genetics and Microbiology, Pontificia Universidad Católica de Chile, Santiago, Chile.

[2]Center for Climate and Resilience Research (CR)2, Chile.

*** Correspondence:**

Beatriz Díez

bdiez@bio.puc.cl

Hot springs, Virus, Lytic, Lysogenic, CRISPR, Viral Metagenomics, Diversity, Cyanophages.

ABSTRACT

Viral infections can vary from lytic (productive) to lysogenic depending on multiple factors such as virus or host genetics, virus host ratio, host physiological state and obviously environmental conditions. Even when lysogenic to lytic viral switch is central to viral ecology, temperate viruses have been much less studied than their lytic counterparts. Inherent difficulty issues with estimating the rates of lysogeny in natural viral communities have obscured the attempts to assess these populations.

Here we used a multi-omic approach along with mitomicyn C in situ inductions, to study lytic and lysogenic viral communities of phototrophic microbial mats from Porcelana hot spring.

Metagenome assembled genomes (MAGs) from Porcelana microbial mats (Northern Patagonia, Chile) were interrogated for the presence of integrated temperate viruses. Metagenomes of natural and mitomicyn C induced viral communities were analyzed to study differential abundance of viral genomes and proteins (functions), as well as ecological parameters such as α and β diversities. Our results suggest that lysogenic and lytic viral populations are strongly associated to specific hosts and consequently to environmental conditions that determines the host community structure. Furthermore,

the most active and dominant bacterial taxa (such as the cyanobacteria *Fischerella*) showed the presence of abundant CRISPR spacers against the most abundant lytic viral populations (cyanophages), while heterotrophic Protobacteria and Firmicutes were associated to spontaneously and mitomycin C induced lysogenic viruses respectively.

This work shows for the first time the different lifestyles of the viral thermophilic communities in phototrophic microbial mats, revealing the nexus between the microbial roles (metabolism) and the type of viral infectious cycle.

1. INTRODUCTION

Prokaryotes and their viruses are present in every imaginable environment on earth, significantly altering the biosphere and their processes as being major players on the nutrient and biogeochemical cycles (Howard-Varona et al., 2017; Suttle, 2007). Viruses are essential components of microbial communities contributing to their fitness and evolution, trough their infective cycle. The lifecycles of viruses have been described to follow three specific paths: lytic (productive), lysogenic and pseudolysogenic (Miller and Day, 2008). Lytic or productive infection are characterized by a rapid replication of phage genome, transcription and translation of viral component to finally release of viral particles by cell lysis (Miller and Day, 2008). Lysogeny and pseudolysogeny are characterized by the persistence of the viral genome inside the host cell but without the production of viral particles and then not producing cell death. These two states, differ in the stable integration of the viral genome into the chromosome of the host, as in pseudolysogeny the viral genome remains unestable, i.e., circularized as a low copy plasmid, and usually failing in its replication as productive infection or when the cell is dividing (Miller and Day, 2008).

The type of viral lifecycle have different effects in the ecology of the host. The lytic infection usually leads to coevolution in an aggressive ‘arms race’ in which viruses and host constantly evolve resistance mechanisms to each other (Rohwer and Thurber, 2009). That is how lytic viruses directly influence the diversity of their hosts by selectively lysing the dominant taxa, which usually are the most active ones in the microbial communities (Suttle, 2007). Consequently, the lytic lifestyle responds to the ecological model of “killing the winner” (Knowles et al., 2016, 2017; Rohwer and Thurber, 2009).

Lysogenic and pseudolysogenic infections in turn, lead to a symbiotic relation with their host, forming a new biological entity known as lysogen(Knowles et al., 2017; Miller and Day, 2008). Lysogenic cycle

brings benefits to hosts, that includes prophage mediated immunity against other virus infections, protection from grazers predation by the acquisition of new virulence factors and gain of new metabolic functions through transduction (Feiner et al., 2015; Howard-Varona et al., 2017; Knowles et al., 2017). Temperate viruses can also influence the microbial communities when they enter in a productive cycle by lysing competitor strains or lysogenizing other microorganisms (Howard-Varona et al., 2017). Also they can produce cooperative effects in the communities by liberating intracellular contents for neighboring cells to use as nutrients(Howard-Varona et al., 2017). Therefore, lysogenic cycle respond to other ecological model where viruses will seek to establish a symbiotic relationship with their host lysogenizing the dominant and most active taxa in the microbial community, following the proposed "piggyback-the-winner" model (Knowles et al., 2016, 2017).

Lytic and lysogenic dynamics have been studied mostly in aquatic environments (marine and freshwater) with a few studies focused in sediments and soils (Howard-Varona et al., 2017; Knowles et al., 2017). Most of thsee studied environments contain complex mixed natural communities which made more complex the study of these dynamics in the field, although the overall conclusion is that lysogeny is a common viral strategy in all the ecosystems (Howard-Varona et al., 2017; Knowles et al., 2017).

Hot springs environments harbor microbial communities dominated by a limited variety of microorganisms and then used as simplified models to understanding how abiotic factors shape the microbial community structure (Inskeep et al., 2010, 2013). Viruses have proven to be ubiquitous, numerous and active components of hot springs microbial communities over the world (Bolduc et al., 2012, 2015; Breitbart et al., 2004; Guajardo-Leiva et al., 2018; Gudbergsdóttir et al., 2016; Menzel et al., 2015; Munson-Mcgee et al., 2018; Schoenfeld et al., 2008; Sharma et al., 2018; Zablocki et al., 2017). Viruses constitute the major biotic factor that can shape the diversity of their host through co-evolution (Guajardo-Leiva et al., 2018; Sano et al., 2018), and regulating the structure of cellular communities through predation (Breitbart et al., 2004; Klatt et al., 2013; Schoenfeld et al., 2008).

Hot springs with circumneutral pH and temperature ranges between 45-70 ºC are dominated by phototrophic microbial mats (Alcamán-Arias et al., 2018; Inskeep et al., 2013; Menzel et al., 2015). Phototrophic microbial mats are a consortia of different microbial groups that are vertical stratified and embedded in an organic matrix on the interface between water and a solid substrate (Bolhuis et al., 2014). The uppermost layer is usually formed by oxygenic phototrophic cyanobacteria from three genera, *Synechococcus* spp. *Oscillatoria* spp., and *Fischerella* spp., and anoxygenic phototrophs

(FAPs), such as Roseiflexus sp. and Chloroflexus sp(Alcamán-Arias et al., 2018; Bhaya et al., 2007; Klatt et al., 2013; Mackenzie et al., 2013; Miller et al., 2006). The deeper layers are usually conformed by a plethora of heterotrophic bacteria and archaea that interact with the primary producers through element and energy cycling (Klatt et al., 2013).

Viral communities that predate in these microbial mats, have been poorly characterized and studies have focused only on the lytic community (Davison et al., 2016; Guajardo-Leiva et al., 2018; Heidelberg et al., 2009). Currently, only one microbial mat viral metagenome constructed from Octopus spring in Yellowstone National Park (YNP) is public available (Davison et al., 2016). Unfortunately this metagenome was constructed using the MDA method, which is known to distort viral abundances in the community (Kim and Bae, 2011), therefore the validity of analyzes based on abundance becomes questionable. Despite this, the comparison of this virome with CRISPR spacers and nucleotide motives frequencies obtained from contigs of the cellular metagenome from the same hot spring showed the presence of viral groups that infected the dominant bacterial genera such as *Synechococcus*, *Roseiflexus*, and *Chloroflexus* (Davison et al., 2016). More recently, viral sequences recovered from metagenomes and metatrancriptomes of a microbial phototrophic mat in a hot spring from Patagonia in Chile, showed that Caudovirales order dominates in this type of microbial consortia (Guajardo-Leiva et al., 2018). In particular cyanophages, viruses that infect cyanobacteria, where one of the most abundant and active Caudovirales in Porcelana mats. This leads to the recovery of the first complete genome of a lytic cyanophage (TC-CHP58) infecting *Fischerella thermalis*, that was part of a monophyletic clade that includes other fresh water non-thermophilic cyanophages that also infected filamentous cyanobacteria (Guajardo-Leiva et al., 2018).

Lysogeny have been historically estimated by quantifying the viral production after temperate virus induction by mitomycin C, that is a DNA damaging agent (Howard-Varona et al., 2017; Kim and Bae, 2018; Knowles et al., 2017). Therefore, this procedure have been usually limited to cultivable bacteria and vary on a strain specific manner. Because of the large number of unknown taxa and strains present in environmental communities the identification and quantification of lysogens in the field remains challenging.

Even though, lysogeny have been considered an effective lifestyle in hot springs (Breitbart et al., 2004; Schoenfeld et al., 2008; Sharma et al., 2018), this strategy have not been systematically studied in these environments, and only assumed by the randomly presence of prophages or molecular markers as

integrases. Breitbart *et al*, in 2004, by epifluorescence microscopy showed for the first time that after experimental mitomycin C inductions there was an increase of 1.2 to 1.4 fold in the VLPs counts from planktonic communities in California hot springs (74-82 °C). Later, on Schoenfeld *et al,* in 2008, viral metagenomes from hot springs were obtained for the first times, finding 86 genes of integrases in two YNP hot springs (74-93 °C), and proposed lysogeny lifestyle as a common feature in thermal aquifers. Recently, Sharma *et al,* in 2018 recovered 66 viral genomes from cellular metagenomes of a Himalayan hot spring (50-98 °C), where 47% of the recovered viral genomes were found to be lysogenic.

Here, in order to unravel the lytic an lysogenic viral communities in thermophilic phototrophic mats we use the circumneutral thermophilic hot spring of Porcelana, to examine the abundance and diversity of natural and mitomycin C (MitC) induced viral comunnities using four viral metagenomes from two Porcelana sites (P50, P55) and 34 bacterial genomes (MAGs) reconstructed from published microbial metagenomes obtained from the same hot spring (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Guajardo-Leiva et al., 2018).

Our analyses of k-mer frequencies, protein clusters (vPCs) and populations (vOTUs) revealed genetic and compositional dissimilarities between viral communities from different temperature sites in the hot spring, and also between natural and experimentally induced communities. The results suggest that lytic lifestyle was predominant in viral populations that infected the most active and abundant phyla (Cyanobacteria and Chloroflexi), but also that lysogenic lifestyle was broadly distributed in viral populations that infected the heterotrophic phyla Proteobacteria and Firmicutes.

2. MATERIAL AND METHODS

Prophage identification in Porcelana metagenome assembled genomes (MAGs).

Assemblies of Porcelana phototrophic mat metagenomes (Alcamán-Arias et al., 2018; Guajardo-Leiva et al., 2018) were taxonomically grouped into Metagenome Assembled Genomes (MAGs) as in (Alcorta et al., 2018; Guajardo-Leiva et al., 2018) with modifications in assembly, quality selection and taxonomic classification steps. Briefly, metagenomes were assembled using De Bruijn graphs implemented in Spades assembler (metaspades) (Bankevich et al., 2012). Later, assemblies for each sites, were taxonomically grouped (bins) using the Expectation–Maximization (EM) algorithm implemented in MaxBin 2.0 (Wu et al., 2016). Completeness and contamination of each bin was

assessed using CheckM (Parks et al., 2015) and taxonomic classification of each bin was accomplished using GTDB taxonomy (Parks et al., 2018) with selection criteria of quality score ≥ 50 (≥ 90% complete, ≤ 10% contaminated) (Parks et al., 2017). A total of 34 MAGs from 9 phyla where recoverd and analyzed for the presence of temperate viruses using PHASTER (Arndt et al., 2016). Only "intact" regions under completeness classification were subsequent analyzed.

Viral enrichment and Mitomycin C (MitC) *in-situ* induction.

Porcelana hot springs (42º 27' 29.1''S-72º 27' 39.3''W) is located in the Chilean Patagonia. Porcelana have a circumneutral pH range of 7.1 to 6.8 and temperatures ranging from 60 ºC to 46 ºC, when sampled on December 2014. Viral communities from phototrophic microbial mats growing in ponds where surface water reached 55 ºC (pH 6.67) or 50 ºC (pH 7.24), hereafter referred as sites P55 and P50 respectively, were sampled at noon (12:00 PM) as described in (Guajardo-Leiva et al., 2018). Briefly, five liters of interstitial fluid was squeezed using 150 µm sterilized polyester net SEFAR PET 1000 (Sefar, Heiden, Switzerland), transported in dark blue PET drums and stored at 4 ºC until filtration. Interstitial fluid was filtered through 0.8 µm pore-size polycarbonate filters (Isopore ATTP, 47 mm diameter, Millipore, Millford, MA, USA) using a Swinex filter holder (Millipore) and 0.22 µm pore-size (Sterivex PES, Millipore). Particles in the 0.22 µm filtrate were concentrated to a final volume of approximately 35 ml using a tangential-flow filtration cartridge (Vivaflow 200, 100 kDa pore size, Vivascience, Lincoln, UK).

Mitomycin C (MitC) experiments were carried *in-situ* in the hot spring. Briefly 500 cm3 of microbial mats from sites P50 and P55 were incubated by separate in polycarbonate transparent bottles, in a total volume of 750 mL of hot spring water from the same site. Mitomycin C (MitC) was added to a final concentration of 1 µg /mL in each bottle and incubated for 24 hours at the same ponds. After incubation, microbial mat and water was squeezed and filtered through a 150 µm sterilized polyester net and transported to the laboratory for serial filtration and concentration as described before.

Purification of viral particles, DNA extractions and high throughput sequencing.

Purification of viral particles was done using CsCl density gradient ultracentrifugation as described in (Thurber et al., 2009). Briefly 8 mL of virus enriched interstitial fluid was brought up to 1.12 g mL-1

with CsCl and then loaded onto the heavier CsCl layers of 1.7, 1.5, 1.35 and 1.2 g mL-1 CsCl prepared using SM buffer (NaCl 100mM, MgSO4 8mM and Tris-Cl 50 mM pH 7.5). Then it was centrifuged at 21,755 RPM for 2 h at 4ºC in a swinging bucket rotor Beckman SW40Ti. DNase treatment of the viral CsCl fraction was used to remove the remaining free DNA from the cellular fraction using 300U of DNAse I for each 1 mL of sample. Viral DNA was extracted as previously described in (Thurber et al., 2009) by formamide/cetyltrimethylammonium bromide (CTAB) method, followed by phenol/chloroform method. Quality and quantity of the extracted nucleic acids were checked and kept at -80 °C. Bacterial DNA contamination was checked by 16S rRNA gene PCR amplification using bacterial universal primer (357F CTCCTACGGGAGGCAGCAG and 907R CCGTCAATTCMTTTRAGTTT) as described in (Mackenzie et al., 2013).

Purified DNA samples were then sequenced by Illumina Mi-seq technology (Roy J. Carver Biotechnology Center, Illinois, USA). Briefly, shotgun viral DNA libraries were prepared with KAPA Hyper Prep kit (Kapa Biosystems, Wilmington, Massachusetts, USA). Libraries were pooled, quantified by qPCR and sequenced on one MiSeq flowcell for 251 cycles from each end of the fragments, using a MiSeq 500-cycle sequencing kit version 3 (Illumina, San Diego. California, USA).

For quality filtering, the following filters were applied using Cutadapt, (Martin, 2011) leaving only sequences longer than 30 bp (-m 30), with a 3' end trimming for bases with a quality below 30 (-q 30), a hard clipping of the first 9 leftmost bases (-u 9), and finally a perfect match of at least 10 bp (-O 10) against KAPA Hyper Prep for Illumina adaptor. Finally, the removal of sequences representing simple repetitions, that are usually due to sequencing errors, was applied using PRINSEQ (Schmieder and Edwards, 2011)DUST threshold 7 (-lc_method dust, -lc_threshold 7). Details of the number of sequences obtained are shown in Supplementary Table S1.

Viral metagenomes assembly and gene prediction.

Viral metagenomes, were assembled using De Bruijn graphs as implemented in the Spades assembler (Bankevich et al., 2012) in metagenomic mode. Only, contigs sequences ≥ 500 pb were further analized.

After assembly, *in-silico* decontamination was performed trough mapping to bacterial, archeal and phiX-174 sequences from RefSeq release 86 using BLASTN (Camacho et al., 2009) (-evalue 0.00001) and query coverage ≥ 5%. Contigs that align to cellular genomes were used to search for temperate viruses using PHASTER (Arndt et al., 2016). Positively identified contigs were returned to the viral contig data sets. Finally, Prodigal software (Hyatt et al., 2010) was used for prediction of protein-coding regions, options (-p meta -n).

Reference dependent taxonomic assignment.

Predicted proteins from contigs dataset were aligned against the NCBI nr database using DIAMOND (Buchfink et al., 2014) (--evalue 0.00001) and parsed using the lowest common ancestor algorithm trough MEGAN 6 (Huson et al., 2016) (LCA score =50) using NCBI taxonomy tree obtaining the taxonomic annotation of each protein.

Abundance of mapped proteins, was quantified through reads recruitment from each viral metagenome using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1). Resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Protein relative abundances were normalized by gene length and library size of each viral metagenome.

***k-mer* frequencies, Viral Protein Clusters and Viral OTUs cataloging as database independent analyses.**

Pairwise mutation distance was calculated for reads that assembly into contigs dataset in all samples using MinHash dimensionality-reduction technique implemented in MASH (Ondov et al., 2016) options (-k 19 and -s 9,999,999).

Predicted proteins ≥ 60 aa in contigs dataset were used to form protein clusters (vPCs) from all viral metagenomes as described on (Brum et al., 2015; Yooseph et al., 2007). Proteins were self clustered by Cd-hit (Li and Godzik, 2006), at 60% of identity and 80% of coverage. After clustering step, all vPCs were quantified in each viral metagenome using Bowtie2 (Langmead and Salzberg, 2012) parameters (-

end-to-end -very sensitive -N 1) and resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Relative abundances were normalized by gene length of each cluster and library size of each viral metagenome as described in (Brum et al., 2015; Yooseph et al., 2007).

vPCs were functionally annotated using Pfam database through hmmscan options (--cut_ga) implemented on HMMER3 (Eddy, 2009) using representative sequences of each vPCs.

vPCs normalized abundance were used to calculate alpha diversity by Shannon H index, evenness by Pielou's index and species richness by using a rarefaction equivalent to the 25% of the total number of vPCs. Also beta diversity, using Bray-Curtis distance was calculated in Vegan package (Oksanen et al 2018).

Viral contigs ≥ 5 kb were used to form viral OTUs (vOTUs) defined here as clustered contigs from all samples at ≥95% identity and ≥ 80% coverage, using nucmer algorithm implemented in MUMmer3 (Kurtz et al., 2004) as in (Duhaime et al., 2017; Paez-espino et al., 2019; Roux et al., 2017). Sequences from temperate viruses, defined as "intact" regions in the PHASTER analyses of Porcelana MAGs, were added to this vOTUs set.

Each vOTU was quantified in each viral metagenome using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1) and resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Only if ≥ 75% of the vOTU sequence length is covered, the vOTU was considered as present in the sample. Relative abundances of viral vOTUs were normalized by Trimmed Mean of M-values (TMM) algorithm, implemented in EdgeR package (Robinson et al., 2009) and vOTU sequence length as in (Roux et al., 2017). Normalized abundance are used to calculate species alpha and beta diversity sing the Vegan package (Oksanen et al., 2018). Alpha diversity was measured by Shannon H index, evenness by Pielou´s index and species richness by using a rarefaction equivalent to the 25% of the total number of vOTUs. Beta diversity was measured through Bray-Curtis distance using Vegan package (Oksanen et al 2018).

Statistical Analysis

MASH distances between samples were visualized by hierarchical clustering (hclust function in R) using minimal increase of sum of squares method (Ward's method). Bray–Curtis distance matrices (vPCs and vOTUs) were visualized by Principal Coordinates Analysis (PCoA) plot, using ampvis2 R package (Albertsen et al., 2015). To explore possible associations between viral community structure, hot springs sites physicochemical properties (pH and temperature) and MitC induction, the vectors of significant hot springs physicochemical factors ($P < 0.05$) and condition were fitted onto PCoA ordination space using the 'envfit' function of the vegan R package (Oksanen et al., 2018) with maximum random permutations.

The rank-based permutation tests of Brunner-Dette-Munk (BDM test, "BDM.2way" function in asbio R package), was used to analyze the influence of MitC induction and sampling site in vPC and vOTU abundance data. Later, a pairwise Wilcoxon test (function pairwise.wilcox.test in R) with bonferroni correction was used to determine which group differences were statistically significant.

Differential abundance of vOTUs (FDR ≤ 0.01) and vPCs (FDR ≤ 0.05) between natural and mitomycin C induced samples were assessed using a paired test in edgeR (Robinson et al., 2009) under phyloseq package (McMurdie and Holmes, 2013, 2014).

Lysogenic viruses proteomic tree.

Genomes of viruses (vOTUs) with differential abundance between induced (MitC) and natural samples, together with temperate viruses found using PHASTER (Arndt et al., 2016) in vOTUs and MAGs datasets, were taxonomically analyzed through a proteomic tree using ViPTree (Nishimura et al., 2017) and reference genomes of dsDNA viruses.

CRISPR spacers and hexamer frequencies for host assignment.

CRISPR loci were identified in MAGs of Porcelana, (Alcorta et al., 2018; Guajardo-Leiva et al., 2018), using CRISPRFinder tool (Grissa et al., 2007). Spacers from CRISPR loci were mapped to vOTUs set using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1).

Porcelana MAGs as well as the vOTUs set from all samples were used for hexamer frequency analysis by VirusHostMatcher (Ahlgren et al., 2017).The virus host pair with the lowest hexamer distance was calculated by d2* and pairs with a distance value ≤ 0.25 (Ahlgren et al., 2017) were used to deduce potential hosts.

Monopartite network analysis of viral communities.

Protein monopartire networks implemented in vContact2 (Jang et al., 2019) iVirus (Bolduc et al., 2017b) tool were constructed using vOTUs set (considered here as viral genomes) from all samples. Briefly, predicted proteins from vOTUs were compared by DIAMOND (Buchfink et al., 2014) in an all-versus-all pairwise comparison (-evalue 0.00001, bitscore 50). Protein clusters were subsequently identified using ClusterONE algorithm (Nepusz et al., 2012) based on DIAMOND e-values with an inflation value of 2, building protein cluster profiles for each genome and generating a similarity network.

For network visualization we used an edge-weighted spring embedded model implemented in Cytoscape 3.7.1(Shannon et al., 2003), which places the genomes sharing more proteins closer to each other, conforming viral clusters. Viral clusters (modules) were then organized according to their predicted host from our previous analysis, and clusterization to reference genomes with ICTV taxonomy on viral RefSeq release 85.

3. RESULTS

Prophage hunting in Porcelana Metagenome Assembled Genomes (MAGs).

In order to investigate the presence of prophage sequences integrated into the bacterial genomes of natural microbial populations from Porcelana hot springs, three previous published metagenomes from the temperature gradient between 48 and 66 °C were assembled and analyzed (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Guajardo-Leiva et al., 2018). The 34 MAGs obtained (≥ 90% genome completeness and ≤ 10% contamination) were distributed between nine different phyla: Acidobacteria (1 MAG), Armatimonadetes (3 MAGs), Bacteroidetes (6 MAGs), Chloroflexi (10 MAGs),

Cyanobaceria (4 MAGs), Deinococcus-Thermus (1 MAG), Planctomycetes (4 MAGs), Proteobacteria (4 MAGs), and Verrucomicrobia (1 MAG).

Prophage analyses of Porcelana MAGs using PHASTER (Table 1) leads to the recovery of a unique complete temperate viral sequence (MAG Nº30) from Burkholderiaceae family (genus GJ-E10) within the Betaproteobacteria class. This prophage sequence (PP_Burkholderia_GJ-E10) was 28.2 Kb length and had a GC content of 68.61%. This region did not present tRNAs but showed 2 integration sites (attL: GTCGGTGAGCAT and attR: GTCGGTGAGCAT) and had 43 predicted proteins among which were found a transposase and an integrase, besides viral hallmark genes such as virion, tail, capsid and portal proteins.

Furthermore, many incomplete prophage (IPP) sequences were also found in several other phyla, Acidobacteria (2 IPPs), Armatimonadetes (3 IPPs), Bacteroidetes (2 IPPs), Chloroflexi (20 IPPs), Cyanobaceria (9 IPPs), Deinococcus-Thermus (3 IPPs), Planctomycetes (6 IPPs), Proteobacteria (6 IPPs), and Verrucomicrobia (1 IPP), most of them lack of integrases, recombination proteins or integration sequences (tRNAs and attachment sites).

Viral metagenomes assemblies.

To understand and analyze the lifestyles of viral communities of Porcelana, four viral metagenomes from natural (P50NAT, P55NAT) and mitomycin C (MitC) induced (P50MitC, P55MitC) viral communities were sequenced. A high number of reads was obtained (3.01 to 8.67 Millions of reads), becoming the viral metagenomes of hot springs with the highest number of sequences available to date. A summary of the obtained sequences is presented in Supplementary Table S1.

Assembly of these viral metagenomes yield from 6197 to 18303 contigs ≥500 pb (Supplementary Table S1), using 35-62% of the quality filtered reads. The number of predicted proteins showed a slight dispersion (13087 to 34233) depending directly on the number of assembled contigs in each sample.

Database dependent analyses through LCA.

Predicted proteins on contigs ≥ 500pb were used for a database dependent taxonomic analyses using NCBI nr, as a first insight to compare and determine differences in taxonomic composition between Porcelana natural and MitC induced viral communities.

Proteins that map to NCBI nr database, represented 13 to 31% of the quality filtered reads in each sample (Supplementary Table S1). The analyses evidenced that in most of the samples, predicted proteins are unknown and absent in current databases. Classification of predicted proteins at Domain level (Figure 1A) showed that 38 to 89% of the mapped reads were assigned to proteins of cellular origin mostly from bacteria, where MitC induced community from site P50 (P50 MitC) exhibit the larger number of these sequences (89%). Moreover, 4 to 59% of the viral metagenomics mapped reads, fell under unclassified sequences category of the NCBI taxonomy tree, with P55 natural community (P55 NAT) showing the larger number of this unclassified sequences (59%).

Proteins classified from viral origin (Figure 1A) fluctuated between 3 to 9% of the total mapped reads. Generally, most of the reads from natural communities (P50 and P55 NATs) mapped to viral proteins of Podoviridae family (60 and 71% respectively). However, Podovirus group present a drastic decrease in MitC induced communities (P50 and P55 MitCs), falling to 31 and 27%, respectively (Figure 1B). On the other hand, most of the reads recovered from MitC induced communities mapped to Myoviridae family (50% on P50 MitC) or to environmental and unclassified viruses (54.5% on P55 MitC), respectively (Figure 1B). Interestingly, viral proteins from Inoviridae family were only present in MitC induced communities (0.5 and 3.2% on P50 MitC and P55 MitC, respectively).

Viral community structure analyses: *k-mer* frequencies, viral Protein Clusters (vPCs) and viral Operational Taxonomic Units (vOTUs).

Contigs ≥ 500pb where also used to catalog the unknown viral sequences space of Porcelana viral communities. A database independent analyses using *k-mer* frequencies, vPCs and vOTUs allowed to generate genetic and compositional dissimilarities metrics to compare and quantify differences between Porcelana natural and MitC induced viral communities.

Community sequences (reads) that assembled into contigs ≥ 500pb were analyzed to estimate genetic distances, using MinHash dimensionality-reduction technique, implemented in MASH. Minimal increase of sum of squares clustering method (Ward's method) based on MASH genetic distance, of *k-mer* frequencies from community reads, showed two clusters of samples (Figure 2A). Both clusters grouped together samples from the same sites (P50 and P55) regardless of induction with MitC. Distance between natural and MitC induced communities (MitC and NAT) inside each cluster was different between sites, with site P50 showing a higher distance (Figure 2A).

Protein clusters (vPCs) of predicted proteins ≥ 60 aa (88979 proteins) on contigs ≥ 500pb were used to analyze and quantify the differences between the protein universe of Porcelana natural and MitC induced viral communities. A total of 9505 of two or more proteins vPCs were obtained, which represent 26 to 52% of the quality filtered reads (Supplementary Table S1). The largest vPC contained 15 proteins, while 66% of the vPCs contained two proteins. Functional annotation of vPCs using Pfam database, showed that 1296 clusters were assigned to 615 protein family models in the database. (Suplementary table S2).

Principal Coordinates Analysis (PCoA) based on Bray-Curtis distance between viral communities from different sites and conditions (Figure 2B), showed a compositional dissimilarity between the two Porcelana sites as well as between natural and MitC induced viral communities. The first two axis of the PCoA explained about 89% of the total variance, where axis 1 separated the communities according mainly to sites (P50 and P55) explaining 59.9% of the total variance, and axis 2 according to the condition (MitC and NAT) explaining 29.3% of the total variance. Ordination showed that dissimilarity between natural and MitC induced community was larger in the P50 than in the P55 site, and also that the MitC induced community at P50 was the most dissimilar. Additionally, vectors of environmental variables were fitted onto ordination space, to explore the effect of pH/temperature and MitC induction, however none of those factors correlated with the viral communities ordination ($P > 0.05$, 9999 permutations).

A rank-based permutation tests (BDM test) was used to analyze how vPC abundance data was affected by the sampling site and MitC induction (Table 2). The test showed that both factors and the interaction between them have a statistically significant ($P > 0.001$) effect on the observed vPCs abundances distribution. Wilcoxon test with bonferroni correction (Table 3) was used to determine which group (site and condition) differences were statistically significant. The latter, found significant differences ($P > 0.001$) between different sites and also between natural and MitC induced viral communities.

Viral OTUs were generated using contigs ≥ 5000 pb (833 contigs) to analyze and quantify the differences between viral populations and genomes from Porcelana natural and MitC induced viral communities. A total of 755 vOTUs were obtained, corresponding to the 25 to 44% of the total quality filtered reads in the samples (Supplementary Table S1).

Principal Coordinates Analysis (PCoA) based on Bray-Curtis distances between viral communities from different sites and condition (Figure 2C), alsoshowed a compositional dissimilarity between the two Porcelana sites as well as between natural and MitC induced viral communities. The first two axis of the PCoA explained about 87% of the total variance, where axis 1 separated the viral communities according to sites (P50 and P55) explaining 55.5% of the total variance, and likewise axis 2 according to the condition (MitC and NAT) explaining 29.3% of the total variance. Once again, ordination showed that dissimilarity between natural and MitC induced communities was larger in the P50 than in the P55 site. Vectors of environmental variables and conditions were as for the vPCs analyses not correlated with the vOTUs based ordination ($P > 0.05$, 9999 permutations).

A rank-based permutation tests (BDM test) was used to analyze how vOTUs abundance data was affected by the sampling site and MitC induction (Table 4). The test showed that sampling site and the interaction between site and MitC induction have a statistically significant ($P > 0.001$) effect on the observed vOTUs abundances distribution. Wilcoxon test with bonferroni correction (Table 5) was used to determine which group (sites and condition) differences were statistically significant. This test, found significant differences ($P > 0.001$) between the natural viral communities from both sites (P50 and P55) with the MitC induced communities, independently of which site they came from.

Cataloging of viral sequences as vPCs and vOTUs provided metrics to explore viral communities diversity, through quantification of these markers in the samples (Supplementary Figure S1). The calculated alpha diversity (Shannon's index) of vPCs and vOTUs (Supplementary Figure S1A) represented an average of 5.47 ± 0.83 and 4.68 ± 0.17, respectively. Two samples were below the average diversity of vPCs and vOTUs, these were the natural community of site P55 (P55NAT) and the MitC induced community of site P50(P50MitC), respectively.

The same pattern was repeated for evenness (Supplementary Figure S1B) and richness (Supplementary Figure S1C) measurements, where P55NAT and P50MitC were below the average evenness of vPCs (0.63 ± 0.09) and vOTUs (0.77 ± 0.03), and also below the average richness of vPCs (646 ±112) and vOTUs (86 ± 5).

Lysogenic viral markers and prophage genomes in Porcelana viral metagenomes.

Analysis based on databases and proteins or genomes differential abundances were conducted to unveil if differences between natural and MitC induced viral communities were related to specific lysogenic markers (vPCs) and prophage genomes (vOTUs).

A manual search was conducted on vPCs with Pfam annotations that have been observed in viruses with a lysogenic lifestyle (Supplementary Figure S2). In addition, an analysis of differential abundance was carried out to determine which vPCs with or without known function were found deferentially abundant on MitC induced and natural samples (Figure 4). Most of the proteins usually found on prophage genomes such as integration and recombination proteins, did not show an association between their abundance and the MitC induction (Supplementary Figure S2)and they even showed higher abundance in the natural communities. A differential abundance analysis was conducted by using Phyloseq (phyloseq_to_edgeR function) discovering that 12 vPCs were statistically more abundant in MitC induced communities than in natural samples, and also that none of them had homology to Pfam protein families related to any known lysogenic core function.

Genome analyses of Porcelana vOTUs using PHASTER found 14 vOTUs that contain lysogenic signatures and where identified as complete prophages, which increases to 15 sequences by including the prophage previously found in the MAG Burkholderia_GJ-E10. In addition, the differential abundance analysis identify 10 vOTUs with statistically significant changes in its abundance between natural and MitC induced viral communities and with two of them also found in the PHASTER analysis. Relative abundances of the 23 sequences found as putative prophages showed that the abundance of these viral sequences was correlated to the MitC induction but in a site dependent fashion (Supplementary Figure S3). That is how at P50 site was possible to observe that 18 vOTUs increased their abundance in the MitC induced community while at P55 site, 13 vOTUs showed this pattern including here the *Burkholderia* prophage.

Finally a viral proteomic tree analyses (Figure 4) was conducted to taxonomically classify in basis to their genomic sequences the 23 putative lysogenic vOTUs found in the previous analyses including the prophage sequence of MAG Burkholderia_GJ-E10. This analysis, revealed 12 groups of genomes mostly related to Myoviridae (11 vOTUs) and Siphoviridae (11vOTUs) families and, one genome related to Podoviridae family. Putative hosts of these 23 vOTUs were mostly from phyla Firmicutes

(61%) and Proteobacteria (35%). Interestingly all vOTUs that were found statistically more abundant in the MitC communities than in the natural communities were classified in Firmicutes infecting viral clades.

Porcelana viral genomes network

A viral protein sharing network analyses was used to clusterize and visualize groups of close viral genomes (genus) from Porcelana natural and MitC induced viral communities.

A total of 775 vOTUs and 2304 reference genomes of archaeal and bacterial viruses (ICTV viral genomes), were used to build a network of 2698 nodes and 168 clusters using vContact2. The final network was formed by 455 vOTUs and 2243 reference genomes (Figure 5), where only 86 Porcelana vOTUs (19%) formed clusters with references genomes, and with most of the Porcelana viruses grouping together into 103 new viral genus (clusters).

Host information was deduced from the 34 Porcelana MAGs by comparison of CRISPR spacers and *hexa-mers* frequencies between the MAGs and vOTUs data sets. Specific clusters were selected from the network, based on host information or clusterization with reference viral genomes to build a sub-network on a more detailed scale (Figure 6) where to discover new pairs of viral genus and their hosts.

This sub-network of 125 cluster was conformed by 455 vOTUs and 520 reference genomes, showing that most of the vOTUs that were part of a viral cluster with reference genomes, corresponded to viruses infecting Protobacteria or Firmicutes. Interestingly, viral clusters infecting Firmicutes were formed mainly by vOTUs recovered from MitC induced communities. On the other hand, most of the vOTUs that were part of clusters (genus) identified by CRISPR spacers, infected *Fischerella*, *Chloroflexus*, *Meiothermus* and *Roseiflexus*. Additionaly, vOTUs that were part of viral clusters identified by hexa-mers frequencies comparisons, infected Pedosphaerales and Burkholderiales. Finally, the identified prophage in Burkholderia GJ-E10 MAG was part of a viral cluster that included 4 other vOTUs recovered from Porcelana natural and MitC induced communities.

Additionally, a viral proteomic tree analysis (Table 6) was made to taxonomically classify the 22 most abundant vOTUs (relative abundance ≥1%). This proteomic analysis showed that 20 vOTUs were classified to viral families inside Caudovirales order, while the two remaining sequences could not be

525 classified into any known group of viruses. The closest viral genomes of reference for the vOTUs were
526 found to infect Actinobacteria (7 vOTUs), Firmicutes (5 vOTUs), Cyanobacteria (4 vOTUS),
527 Methanobacteria (3 vOTUs) and Proteobacteria (1 vOTU). The most abundant genome was the vOTU
528 P55_C_19, which has a predicted lysogenic lifestyle (Supplementary Figure S3) and their closest
529 genome in the vOTU P50MIT_C_17 (Figure 4) followed by the reference Streptococcus phage phiD12
530 (Table 6).
531 Information of the natural hosts in Porcelana inferred by CRISPR spacers was usually not congruent
532 with the host of the closest viral genome or even with hosts of the viral genomes that were part of the
533 clade. Only Cyanobacteria infecting vOTUs (4) had congruent information about the information of
534 natural host in Porcelana and the closest genome in the database (Table 6), showing that these four
535 vOTUs represent an abundant new viral genus (5.2%), with a lytic lifestyle that infected *Fischerella* sp.
536 Two other vOTUs have a match to CRISPRs spacers from *Fischerella* sp., but the closest reference
537 genome infects Archaea of the phylum Methanobacteria (Table 6). Finally, one vOTU matched to
538 CRISPR spacers from *Meiothermus* sp., but its closest reference was infecting the phylum,
539 Actinobacteria.
540
541 **4. DISCUSSION**
542
543 In contrast to the numerous studies of viral communities that inhabit thermal waters of high
544 temperature or acidic pH, the study of viral communities of thermophilic phototrophic microbial mats
545 have been neglected. To our knowledge, phototrophic mats in hot springs have never been studied in
546 relation to the lifestyles of their viral communities using viral metagenomics.
547 Here our database-independent analyses of natural and MitC induced viral communities revealed
548 genetic and compositional dissimilarities between viral lifestyles over different temperatures into the
549 same hot spring. Database and differential abundance analyses to protein (vPCs) and genome (vOTUs)
550 levels, showed that specific groups of viruses associated with specific hosts have different lifestyles and
551 susceptibility to MitC. This was emphasized by the networks analysis, denoting specific viral genus
552 (clusters) formed exclusively by vOTUs recovered from MitC communities while other viral genera
553 were formed by genomes recovered from both natural and induced communities. Together, our results
554 suggest that lytic lifestyle was predominant in viral populations that infected most active and abundant
555 primary producers (Cyanobacteria and Chloroflexi) in the mats, but also that lysogenic lifestyle was

broadly distributed in viral populations that infected heterotrophic Proteobacteria and Firmicutes. Temperate virus infecting Proteobacteria were spontaneously induced by an unknown stimulus, while the Firmicutes infecting viruses were activated by the MitC induction.

Lysogens in Porcelana phototrophic microbial mats.

Temperate bacterial viruses (prophages) can enter in a symbiosis with their host organism, forming a new unit called a lysogen(Knowles et al., 2017). Lysogeny brings benefits to hosts, that includes prophage mediated immunity against other virus infections, protection from grazers predation by the acquisition of new virulence factors and gain of new metabolic functions (Knowles et al., 2017).

It has been argued that lysogens have a high frequency (20 to 60%) within bacterial cultured strains and environmental isolates, even when reported numbers vary widely (Miller and Day, 2008). Also, 66% of the viral genomes in reference databases have been reported to have a lysogenic lifestyle (McNair et al., 2012).

Lysogeny has been historically estimated by quantifying the viral progeny after prophage induction by mitomycin C induction, a DNA damaging agent (Howard-Varona et al., 2017; Kim and Bae, 2018; Knowles et al., 2017). However, this methodology of induction has been mostly used in cultivable bacteria and its effects in natural communities are not well understood (Kim and Bae, 2018; Knowles et al., 2017). Hence, the identification and quantification of lysogens in mixed natural communities have been even more challenging, than in pure cultures.

In order to solve culture limitations and study the lysogens present in the natural microbial community of Porcelana, a genome-resolved metagenomic analysis (MAGs) was used. We used 34 MAGs recovered from three previously published cellular metagenomes of Porcelana (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Guajardo-Leiva et al., 2018) that meet the completeness and contamination standards (≥ 90% genome completeness and ≤ 10% contamination) (Bowers et al., 2017) developed by the Genomic Standards Consortium (GSC) and that were above the threshold (>70% completeness) used in a recent study of lysogens populations from the murine gut microbiota (Kim and Bae, 2018). Besides the high quality of the bacterial genomes recovered, the presence of active lysogens was low. We discovered one unique temperate virus in MAG Nº30 (Burkholderia GJ-E10), whereas most of the viral sequences found corresponded to non-active vestigial prophages (Table 1). These vestigial sequences, lack known lysogeny markers such integration enzymes, recombination enzymes, tRNAs

587 and attachment sites (Canchaya et al., 2003). However, these defective prophage sequences have been
588 reported to provide adaptive functions to bacteria such as, gene transfer agents (GTAs) that
589 scholastically transfer fragments of chromosomal DNA to other microorganism, also type 6 secretion
590 systems (T6SSs) and bacteriocins who are involved in bacterial defense and competition mechanisms
591 (Bobay et al., 2014).
592
593 Nevertheless, these genome-resolved metagenomic analyses have a possible bias given the difficulty of
594 assembling genomes from populations below 1% relative abundance (Albertsen et al., 2013).
595 Therefore, it will be difficult to find the lysogens of those low abundant taxa, and their information will
596 be neglected.
597
598 **Database dependent taxomic annotation of natural and MitC induced viral communities.**
599
600 The database dependent analyses (Figure 1A), showed that a high number of reads in the viral
601 metagenomes were assigned to proteins from cellular origin especially to Bacteria which is common
602 feature in viral metagenomics studies (Edwards and Rohwer, 2005; Roux et al., 2012). This bias relies
603 in the lack of viral gene annotation in databases, and is also explained by the HGT between viral and
604 host genomes, that leads to incorrect annotation based on the closest homologous sequences (Edwards
605 and Rohwer, 2005; Roux et al., 2012). This misleading annotation is especially evident in
606 underexplored communities such as those from hot spring environments, since only 17 representatives
607 of thermophilic bacterial viruses exist in the databases (Zablocki et al., 2018). In this study, the
608 increase in the proportion of viral sequences assigned to bacteria in communities induced by MitC,
609 compared to natural communities, is also due to the fact that the prophages sequences in the databases
610 are annotated within the taxonomic domain of the lysogen (Canchaya et al., 2003) that in this case is
611 Bacteria.
612 Viral proteins assigned to viruses separated here the natural from the MitC induced viral communities
613 (Figure 1B), whereas both communities were dominated by Caudovirales order. Natural communities
614 were rich in viruses from Podoviridae family (60-71%),as previously reported in the same hot spring
615 (Guajardo-Leiva et al., 2018) as well as in Brandvlei hot spring (South Africa) (Zablocki et al., 2017,
616 2018), where both springs harbor extended phototrophic microbial mats.

MitC induced communities were in turn dominated by Myoviridae family viruses as well as by environmental or unclassified viruses. Myoviridae family includes known lysogenic lifestyle genera such as Mu-like viruses (*Muvirus*) and some P2-like viruses (*P2virus*) (Canchaya et al., 2003) that can explain the high presence of this family in the induced community.

Diversity variations of natural and MitC induced viral communities at genetic and ecological levels.

Given the potential bias annotation by using database dependent analysis, we cataloged the viral sequences using a database independent aproach, through k-mer frequencies, vPCs and vOTUs unveiling genetic and compositional differences between natural and MitC induced communities but more importantly at different sites in Porcelana (Figure 2). At genetic level (k-mer frequencies) hierarchical clustering showed that communities from the same site (natural or induced) cluster together (Figure 2A). Likewise, at compositional level, PCoA analyses based on Bray-Curtis distance matrix of vPCs (Figure 2B) and vOTUs (Figure 2C) separated communities by site in the first axis, which explained most of the variance (60% and 56%, respectively). This separation of viral communities within a unique hot springs have been reported before for archaeal viruses that inhabit acidic hot springs from YNP (Bolduc et al., 2012; Snyder et al., 2007) and, at the Himalayan hot spring Manikaran (Sharma et al., 2018). A previous study in Porcelana using viral sequences recovered from cellular metagenomes showed the same pattern, where viral communities from sites at different temperatures differ according to changes in the composition of their host community across the thermal gradient (Guajardo-Leiva et al., 2018).

In the other hand, separation of natural and MitC induced viral communities at genetic level was comparatively higher at the P50 site than at the P55 site (distance in cladogram of Figure 2A), similar to results of vPCs and vOTUs compositional analyses (Figures 2A and 2C). It is well known that mitomycin have a taxa specific effect (Miller and Day, 2008; Paul and Kellogg, 2000), even within strains of the same species (Knowles et al., 2017). Since the composition of Porcelana microbial communities changes according to the environmental gradients between different sites in the hot spring (Alcamán-Arias et al., 2018; Guajardo-Leiva et al., 2018) it is expected that the effect of mitomycin will be differential as well. Moreover, it is expected that not all taxa harbor prophages, as reported in a mitomycin induction experiment in fjord waters at British Columbia. There, 80% of the heterotrophic

bacteria and only 0.6% of the cyanobacterial cells harbored prophages (Ortmann et al., 2002), similar result was reported in murine gut microbiota, where most of the prophages sequences found, were associated to Proteobacteria and Firmicutes hosts (Kim and Bae, 2018), suggesting that not all taxa present in a natural environment will carry temperate viruses. In our analyses lysogenic viruses were also associated to Proteobacteria and Firmicutes, independently of the hot spring site.

Statistical tests (BDM and Wilcoxon) confirmed the significance of the dissimilarities found in the vPCs abundances (Table 2 and Table 3) between viral communities of different sites and also between natural and induced communities. For vOTUs only the dissimilarities found between sites and the combination of sites and MitC induction were statistically significant (Table 4 and 5). This result is not strange if we consider that both markers (vOTUs and vPCs) represent different approaches to viral communities. It is a known issue, that vOTUs approximation usually represent the most abundant viral populations, because metagenomic assemblies only yield large contigs for these abundant viral genotypes (Roux et al., 2017), while vPCs by using smaller contigs are able to capture the less abundant genotypes in the community. It is then possible to infer that significant changes in MitC induced viral community were here strongly related to changes in the abundances of the rare viral populations.

Also worth to note that the changes in alpha diversity detected between natural and MitC induced communities, were dependent on the sampling site (Suplementary Figure S1). Diversity and evenness of the MitC induced community measured by vPCs metrics increases at the P50 site, but decreased when vOTUs metrics were used, this can be explained by a greater susceptibility of cellular community here to MitC. This effect generate a profound shift in the viral community, that was completely different from the natural community, being less diverse in terms of richness and evenness, dominated by a lower number of viral genotypes. Likewise, the increase in the alpha-diversity (Shannon index) and their components (evenness and richness) for vPCs and vOTUs metrics in the induced community at site P55, implies that MitC induced new viral genomes and new protein families that were absent in the natural community, and also that the resulting community was more evenly distributed.

Active lysogenic viruses are frequent in natural viral communities of Porcelana.

The search of lysogenic markers have been used as a strategy to study and discover lysogens and prophages in natural communities and isolated bacteria (Howard-Varona et al., 2017). Genes such as integrases and ParA/B (Emerson et al., 2012) or integrases and cI-type repressors (McDaniel et al., 2008) have been used in hypersaline and marine ecosystems respectively, to study lysogenic seasonal dynamics. Besides, databases from lysogenic markers and common viral genes have been used more widely to identify prophages in microbial genomes and metagenomic data (Arndt et al., 2016; Reis-Cunha et al., 2017; Zhou et al., 2011).

In Porcelana lysogenic markers genes (Supplementary Table S2) and other genes present in lysogenic viruses (Supplementary Figure S2) were annotated and commonly found in both natural and MitC induced communities. Moreover, some of these gene markers (e.g. integrases, ParB and recombinases) were found to be more abundant in natural than in the MitC induced communities (Suplementary Figure S2), suggesting once more that lysogeny is a common feature in natural microbial communities of hot springs (Breitbart et al., 2004; Schoenfeld et al., 2008; Sharma et al., 2018). However, it has only been established experimentally in the single work of Breitbart et al, in 2004, which showed that after an in-situ mitomycin C induction in planktonic microbial communities from California hot springs (74-82 °C) there was an increase of 1.2 to 1.4 fold in the number of VLPs detected by epifluorescence microscopy. Therefore, our findings emphasize that lysogeny is a common feature in natural microbial communities of hot springs (Breitbart et al., 2004; Schoenfeld et al., 2008; Sharma et al., 2018).

Additionally the differential abundance analyses implemented in Phyloseq (McMurdie and Holmes, 2013, 2014) allowed us to detect 12 vPCs that showed statistical significant differences in the MitC induced communities. Functional annotations of this vPCs did not correspond to any lysogenic markers (Supplementary Figure S2). However, two interesting protein families arise, FtsK/SpoIIIE family (PF01580) and Tubulin/Fts family (PF00091). Those two families were previously described in lytic and lysogenic viruses infecting *Bacillus* (Grose et al., 2014) or *Clostridium* and *Pseudomonas* (Kraemer et al., 2012) respectively. The FtsK/SpoIIIE family is involved in the control of *Bacillus* host transition into the sporulation state (Grose et al., 2014), contributing to the host environmental fitness. The Tubulin/FtsZ family have two different functions, the partitioning of the non-integrated prophage

genomes during *Clostridium* cell division (Oliva et al., 2012) and the movement of replicating phage chromosomes to the center of the cell to form an efficient phage factory that improves the burst size (Kraemer et al., 2012).

This differential abundance analyses suggest that MitC induction has a strong and quantitative effect over a specific group of hosts, those who in turn harbor specific prophages that contain new unknown functions (vPCs) that have not been described in host springs before.

Additionally in the genomic context, marker based (PHASTER) and differential abundance analyses of Porcelana vOTUs revealed 23 putative lysogenic viruses, which also includes the prophage PP_Burkholderia_GJ-E10 (Supplementary Figure S3). These 23 vOTUs would mostly infect bacteria of the Firmicutes (14) and Proteobacteria (8) phyla, whereas only Firmicutes viruses showed an increment of their relative abundances after mitomycin induction of Porcelana microbial mat at both sites (P50 and P55). Interestingly these two phyla harbor the bast majority of prophage sequences in databases (Canchaya et al., 2003) what has also been described in oceans (McDaniel et al., 2008) and hot springs (Sharma et al., 2018). Metagenomic survey of marine phage integrases showed that a large number of phage integrases close to viruses of Protobacteria (*Vibrio*) and Firmicutes (*Clostridium*) were widespread in GOS samples and also in MitC inductions in Tampa Bay (McDaniel et al., 2008). Morover, metagenomic recovery of viral genomes in a Himalayan hot spring showed that Proteobacteria and Firmicutes phages composed 28 of the 31 lysogenic viruses found in the microbial mats and sediments.

In Porcelana viral communities, lysogenic viruses represented a small fraction of the total number of genomes recovered (vOTUs), and most of them were representatives of less abundant genomes (< 1 % of abundance) (Table 6). However, there was one exception (vOTU P55_C_19) that corresponded to the most abundant (3%) virus in Porcelana (Table 6), with the highest presence in the natural community at P50 site and in the MitC induced community from P55 site. This vOTU shared 10% of their protein sequences with viruses infecting Firmicutes of the *Streptococcus* and *Clostridium* genera.

Together, our results were in accordance with previous studies in hot springs and other environments pointing Proteobacteria and Firmicutes as most common putative hosts of lysogenic viruses in environmental communities. However, temperate phages richness in Porcelana seems to be lower than in the recent report of viruses from Himalayan hot springs (Sharma et al., 2018). This can be explained due to the fact that the dominant bacterial groups in that hot spring are exactly Proteobacteria and

Firmicutes phyla (Sharma et al., 2018), and although in Porcelana these groups are also important, they are far from being the dominant ones (Alcamán-Arias et al., 2018; Guajardo-Leiva et al., 2018).
Our results also suggest that Firmicutes lysogens were more susceptible to MitC induction than Proteobacteria members, and also that Proteobacteria lysogens together with a minor number of Firmicutes lysogens were spontaneously induced in Porcelana hot springs. It is important to emphasize that only these two groups were associated with lysogenic viruses regardless of the induction factor that leads to the develop of a lytic cycle.

Lytic and lysogenic networks reveal new and ecological relevant viral genera in Porcelana communities.

Monopartite protein shared networks implemented on vContact2 (Jang et al., 2019) are able to group together different viral genomes (vOTUs) into Viral Clusters (VCs) which can be considered a genus with an 80% of precision (Bolduc et al., 2017). This methodology allowed us to explore and visualize the existing taxonomic relationships between viral genomes, which is especially relevant in undersampled environments that lack references in databases such as hot springs.

Using this methodology we constructed a network of Porcelana viral communities and RefSeq-ICTV reference viral genomes (Figure 5) that showed a higher degree of modularity (3.6 fold), when compared to the network formed by only the RefSeq (1964) reference genomes (Bolduc et al., 2017). This larger modularity occurs due to the fact that most of the Porcelana genomes (369) grouped together forming 103 new viral genera. This high rate of new genera discovery here, is even higher than that reported for the marine environment (Jang et al., 2019). This is evident when we compare Porcelana results to those of the Global Ocean Sampling campaign (GOS), where the latter had 15280 viral genomes and only 919 represented new genera (Jang et al., 2019).
Most of these new genera discovered in Porcelana will probably remain completely unknown, because most of them (94) lack any information related to taxonomy or host. From these new genera, we identified nine infecting *Fischerella* (2 genus), *Chloroflexus* (1 genus), *Roseiflexus* (1genus), *Meiothermus* (4 genus), *Burkholderia* (1 genus) and *Pedosphaerales* (1 genus) (Chapter 2, Figure 6). The first four new viral taxa infect three genera of bacteria that are known to play an important function in Porcelana microbial mats, as primary producers and builders of these communities (Alcamán-Arias

et al., 2018; Alcamán et al., 2015). Therefore, it is possible that these specific host virus relationships are relevant to the entire microbial community. This is particularly true in the case of *Fischerella*, on which the rest of the community depends, due to its important ecological role as carbon and nitrogen fixer in this and many other hot springs over the world, since two of the new genera that infect this organism are among the most abundant viral genomes recovered from Porcelana (Table 6). The first *Fischerella* infecting viral genus, with 5% of total abundance, belongs to the Podoviridae family, with one full representative genome (TC-CHP58) that was recovered from Porcelana cellular metagenomes (Guajardo-Leiva et al., 2018). The second viral genus belongs to Siphoviridae family and represent ~ 2% of the total viral community. However, the genomes of this new genus do not resemble any other known cyanophage. In that sense, our results emphasize the need to do more extensive work in this type of environments, in order to recover a major number of complete genomes of these and other viruses infecting key organisms, such as those associated to the Chloroflexi phylum, in this type of worldwide distributed hot springs.

Together, the present work represent the firsts analysis of viral metagenomes from induction experiments in hot springs, and also the first coupling genome-resolved metagenomic analysis to viral metagenomes and differential abundance analyses to better understand the viral lifestyles in these extreme environments. Here we demonstrate that lytic lifestyle was predominant in the most abundant viral genera some of which infected one of most active and abundant microorganisms (*Fischerella*) in phototropic mats of circumneutral pH hot springs. Also we corroborate that in hot springs as in many other environments lysogeny is broadly distributed in viral populations that infected Proteobacteria and Firmicutes, where Proteobacteria lysogens were spontaneously induced by still unknown factors, while Firmicutes lysogens were induced by mechanisms that were activated by MitC.

From this, it is possible to propose that the ecological model “kill the winner” is the one that best fits the viral communities of the active primary producers, while “piggyback the winner” adjust better to viral communities of the heterotrophic bacteria that take advantage of the primary production in themophilic phototrophic microbial mats.

CONFLICT OF INTEREST

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

AUTHOR CONTRIBUTIONS

SGL and BD conceived and designed the experiments. SGL performed the experiments. SGL, OS, and BD analysed the data. SGL, and BD wrote the paper. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was financially supported by PhD scholarships CONICYT N° 21130667, 21172022 and CONICYT grant FONDECYT N°1150171.

ACKNOWLEDGMENTS

We are grateful to Huinay Scientific Field Station for making our work in the Porcelana hot spring possible.

REFERENCES

Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). Alignment-free d2∗ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* 45, 39–53. doi:10.1093/nar/gkw1002.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi:10.1038/nbt.2579.

Albertsen, M., Karst, S. M., Ziegler, A. S., Kirkegaard, R. H., and Nielsen, P. H. (2015). Back to basics The influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS One* 10, 1–15. doi:10.1371/journal.pone.0132783.

827 Alcamán-Arias, M. E., Pedrós-Alió, C., Tamames, J., Fernández, C., Pérez-Pantoja, D., Vásquez, M., et
828 al. (2018). Diurnal changes in active carbon and nitrogen pathways along the temperature gradient
829 in porcelana hot spring microbial mat. *Front. Microbiol.* 9, 1–17. doi:10.3389/fmicb.2018.02353.

830 Alcorta, J., Espinoza, S., Viver, T., Alcamán-Arias, M. E., Trefault, N., Rosselló-Móra, R., et al. (2018).
831 Temperature modulates Fischerella thermalis ecotypes in Porcelana Hot Spring. *Syst. Appl.*
832 *Microbiol.* 41, 531–543. doi:10.1016/j.syapm.2018.05.006.

833 Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster
834 version of the PHAST phage search tool. *Nucleic Acids Res.* 44, W16–W21.
835 doi:10.1093/nar/gkw387.

836 Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012).
837 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J.*
838 *Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.

839 Bhaya, D., Grossman, A. R., Steunou, A. S., Khuri, N., Cohan, F. M., Hamamura, N., et al. (2007).
840 Population level functional diversity in a microbial community revealed by comparative genomic
841 and metagenomic analyses. *ISME J.* 1, 703–713. doi:10.1038/ismej.2007.46.

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

844 Bolduc, B., Jang, H. Bin, Doulcier, G., You, Z.-Q., Roux, S., and Sullivan, M. B. (2017a). vConTACT:
845 an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* 5,
846 e3243. doi:10.7717/peerj.3243.

847 Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F., and Young, M. (2012).
848 Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-
849 Dominated Yellowstone Hot Springs. *J. Virol.* 86, 5562–5573. doi:10.1128/JVI.07196-11.

850 Bolduc, B., Wirth, J. F., Mazurie, A., and Young, M. J. (2015). Viral assemblage composition in
851 Yellowstone acidic hot springs assessed by network analysis. *ISME J.* 9, 2162–2177.
852 doi:10.1038/ismej.2015.28.

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L., and Sullivan, M. B. (2017b). IVirus: Facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J.* 11, 7–14. doi:10.1038/ismej.2016.89.

Bolhuis, H., Cretoiu, M. S., and Stal, L. J. (2014). Molecular ecology of microbial mats. *FEMS Microbiol. Ecol.* 90, 335–350. doi:10.1111/1574-6941.12408.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi:10.1038/nbt.3893.

Breitbart, M., Wegley, L., Leeds, S., Rohwer, F., and Schoenfeld, T. (2004). Phage Community Dynamics in Hot Springs These include : Phage Community Dynamics in Hot Springs. *Appl. Environ. Microbiol.* 70, 1633–1640. doi:10.1128/AEM.70.3.1633.

Brum, J. R., Sullivan, M. B., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., et al. (2015). Patterns and ecological drivers of ocean viral communities. *Science (80-.).* 348, 1261498-1–11. doi:10.1126/science.1261498.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 1–9. doi:10.1186/1471-2105-10-421.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. *Microbiol. Mol. Biol. Rev.* 67, 238–76, table of contents. doi:10.1128/MMBR.67.2.238.

Davison, M., Treangen, T. J., Koren, S., Pop, M., and Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS One* 11, 1–23. doi:10.1371/journal.pone.0160574.

Duhaime, M. B., Solonenko, N., Roux, S., Verberkmoes, N. C., Wichels, A., and Sullivan, M. B. (2017). Comparative omics and trait analyses of marine Pseudoalteromonas phages advance the phage OTU concept. *Front. Microbiol.* 8, 1–16. doi:10.3389/fmicb.2017.01241.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. in *Genome Informatics 2009*, 205–211. doi:10.1142/9781848165632_0019.

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510. doi:10.1038/nrmicro1163.

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B., and Banfielda, J. F. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. *Appl. Environ. Microbiol.* 78, 6309–6320. doi:10.1128/AEM.01212-12.

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., and Herskovits, A. A. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol.* 13, 641–650. doi:10.1038/nrmicro3527.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, 52–57. doi:10.1093/nar/gkm360.

Grose, J. H., Jensen, G. L., Burnett, S. H., and Breakwell, D. P. (2014). Genomic comparison of 93 Bacillus phages reveals 12 clusters, 14 singletons and remarkable diversity. *BMC Genomics* 15. doi:10.1186/1471-2164-15-1184.

Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F., and Díez, B. (2018). Active crossfire between cyanobacteria and cyanophages in phototrophic mat communities within hot springs. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.02039.

Gudbergsdóttir, S. R., Menzel, P., Krogh, A., Young, M., and Peng, X. (2016). Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. *Environ. Microbiol.* 18, 863–874. doi:10.1111/1462-2920.13079.

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T., and 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. doi:10.1371/journal.pone.0004169.

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

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* 12, 1–12. doi:10.1371/journal.pcbi.1004957.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11. doi:10.1186/1471-2105-11-119.

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., and Rusch, D. B. (2013). The YNP metagenome project: Environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. *Front. Microbiol.* 4, 1–15. doi:10.3389/fmicb.2013.00067.

Inskeep, W. P., Rusch, D. B., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Richardson, T. H., et al. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* 5. doi:10.1371/journal.pone.0009773.

Jang, H. Bin, Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene- sharing networks. *Nat. Biotechnol.* doi:10.1038/s41587-019-0100-8.

Kim, K.-H., and Bae, J.-W. (2011). Amplification Methods Bias Metagenomic Libraries of Uncultured Single-Stranded and Double-Stranded DNA Viruses. *Appl. Environ. Microbiol.* 77, 7663–7668. doi:10.1128/aem.00289-11.

Kim, M. S., and Bae, J. W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* 12, 1127–1141. doi:10.1038/s41396-018-0061-9.

Klatt, C. G., Inskeep, W. P., Herrgard, M. J., Jay, Z. J., Rusch, D. B., Tringe, S. G., et al. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. *Front. Microbiol.* 4, 1–23. doi:10.3389/fmicb.2013.00106.

933 Knowles, B., Bailey, B., Boling, L., Breitbart, M., Cobián-Güemes, A., del Campo, J., et al. (2017).
934 Variability and host density independence in inductions-based estimates of environmental
935 lysogeny. *Nat. Microbiol.* 2, 17064. doi:10.1038/nmicrobiol.2017.64.

936 Knowles, B., Silveira, C. B., Bailey, B. A., Barott, K., Cantu, V. A., Cobian-Guëmes, A. G., et al.
937 (2016). Lytic to temperate switching of viral communities. *Nature* 531, 466–470.
938 doi:10.1038/nature17193.

939 Kraemer, J. A., Erb, M. L., Waddling, C. A., Montabana, E. A., Zehr, E. A., Wang, H., et al. (2012). A
940 phage tubulin assembles dynamic filaments by an atypical mechanism to center viral DNA within
941 the host cell. *Cell* 149, 1488–1499. doi:10.1016/j.cell.2012.04.034.

942 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004).
943 Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi:10.1186/gb-
944 2004-5-2-r12.

945 Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9,
946 357–9. doi:10.1038/nmeth.1923.

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

949 Mackenzie, R., Pedrós-Alió, C., and Díez, B. (2013). Bacterial composition of microbial mats in hot
950 springs in Northern Patagonia: Variations with seasons and temperature. *Extremophiles* 17, 123–
951 136. doi:10.1007/s00792-012-0499-z.

952 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
953 *EMBnet.journal* 17, 10. doi:10.14806/ej.17.1.200.

954 McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., et al. (2008).
955 Metagenomic analysis of lysogeny in Tampa Bay: Implications for prophage gene expression.
956 *PLoS One* 3. doi:10.1371/journal.pone.0003263.

957 McMurdie, P. J., and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis
958 and Graphics of Microbiome Census Data. *PLoS One* 8. doi:10.1371/journal.pone.0061217.

959 McMurdie, P. J., and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is
960 Inadmissible. *PLoS Comput. Biol.* 10. doi:10.1371/journal.pcbi.1003531.

McNair, K., Bailey, B. A., and Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* 28, 614–618. doi:10.1093/bioinformatics/bts014.

Menzel, P., Gudbergsdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., et al. (2015). Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs. *Microb. Ecol.* 70, 411–424. doi:10.1007/s00248-015-0576-9.

Miller, S. R., Purugganan, M. D., and Curtis, S. E. (2006). Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium Mastigocladus laminosus. *Appl. Environ. Microbiol.* 72, 2793–2800. doi:10.1128/AEM.72.4.2793-2800.2006.

Miller, R. V., and Day, M. J. (2008). "Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology," in *Bacteriophage Ecology*, ed. S. T. Abedon (Cambridge: Cambridge University Press), 114–144. doi:10.1017/CBO9780511541483.008.

Munson-Mcgee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J., Weitz, J. S., et al. (2018). A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments. *ISME J.* 12, 1706–1714. doi:10.1038/s41396-018-0071-7.

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* 9, 471–472. doi:10.1038/nmeth.1938.

Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017). ViPTree: The viral proteomic tree server. *Bioinformatics* 33, 2379–2380. doi:10.1093/bioinformatics/btx157.

Oksanen J, Blanchet G, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens H, Szoecs E and Wagner H. (2019). vegan: Community Ecology Package. R package version 2.5-5. https://CRAN.R-project.org/package=vegan

Oliva, M. A., Martin-Galiano, A. J., Sakaguchi, Y., and Andreu, J. M. (2012). Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci.* 109, 7711–7716. doi:10.1073/pnas.1121546109.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 1–14. doi:10.1186/s13059-016-0997-x.

Ortmann, A. C., Lawrence, J. E., and Suttle, C. A. (2002). Lysogeny and lyric viral production during a bloom of the cyanobacterium Synechococcus spp. *Microb. Ecol.* 43, 225–231. doi:10.1007/s00248-001-1058-9.

Paez-espino, D., Roux, S., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2019). IMG / VR v . 2 . 0 : an integrated data management and analysis system for cultivated and environmental viral genomes. 47, 678–686. doi:10.1093/nar/gky1127.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P. A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi:10.1038/s41564-017-0012-7.

Parks, D. H., Waite, D. W., Skarshewski, A., Chuvochina, M., Rinke, C., Hugenholtz, P., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36. doi:10.1038/nbt.4229.

Paul, J. H., and Kellogg, C. A. (2000). "Ecology of Bacteriophages in Nature," in *Viral Ecology* (Elsevier), 211–246. doi:10.1016/B978-012362675-2/50006-9.

Reis-Cunha, J. L., Bartholomeu, D. C., Earl, A. M., Birren, B. W., and Cerqueira, G. C. (2017). ProphET, Prophage Estimation Tool: a standalone prophage sequence prediction tool with self-updating reference database. *bioRxiv*, 176750. doi:10.1101/176750.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Rohwer, F., and Thurber, R. V. (2009). Viruses manipulate the marine environment. *Nature* 459, 207–212. doi:10.1038/nature08060.

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., and Sullivan, M. B. (2017). Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5, e3817. doi:10.7717/peerj.3817.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 7. doi:10.1371/journal.pone.0033641.

Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. (2018). Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. *Nat. Ecol. Evol.* 2, 510–519. doi:10.1038/s41559-017-0435-9.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi:10.1093/bioinformatics/btr026.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., and Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Appl. Environ. Microbiol.* 74, 4164–4174. doi:10.1128/AEM.02598-07.

Shannon, P., Markiel, A., Owen Ozier, 2, Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 2498–2504. doi:10.1101/gr.1239303.metabolite.

Sharma, A., Schmidt, M., Kiesel, B., Mahato, N. K., Cralle, L., Singh, Y., et al. (2018). Bacterial and Archaeal Viruses of Himalayan Hot Springs at Manikaran Modulate Host Genomes. *Front. Microbiol.* 9, 1–15. doi:10.3389/fmicb.2018.03095.

Snyder, J. C., Wiedenheft, B., Lavin, M., Roberto, F. F., Spuhler, J., Ortmann, A. C., et al. (2007). Virus movement maintains local virus population diversity. *Proc. Natl. Acad. Sci.* 104, 19102–19107. doi:10.1073/pnas.0709445104.

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

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

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi:10.1093/bioinformatics/btv638.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* 5, 0432–0466. doi:10.1371/journal.pbio.0050016.

Zablocki, O., van Zyl, L. J., Kirby, B., and Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. *Viruses* 9. doi:10.3390/v9110348.

Zablocki, O., van Zyl, L., and Trindade, M. (2018). Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages. *Extremophiles* 22, 827–837. doi:10.1007/s00792-018-1052-5.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. *Nucleic Acids Res.* 39, 347–352. doi:10.1093/nar/gkr485.

Data Availability Statement

The datasets generated for this study can be found NCBI as follow: Access to raw data for metagenomes and metatranscriptomes is available through NCBI BioProject ID PRJNAxxxx

FIGURES

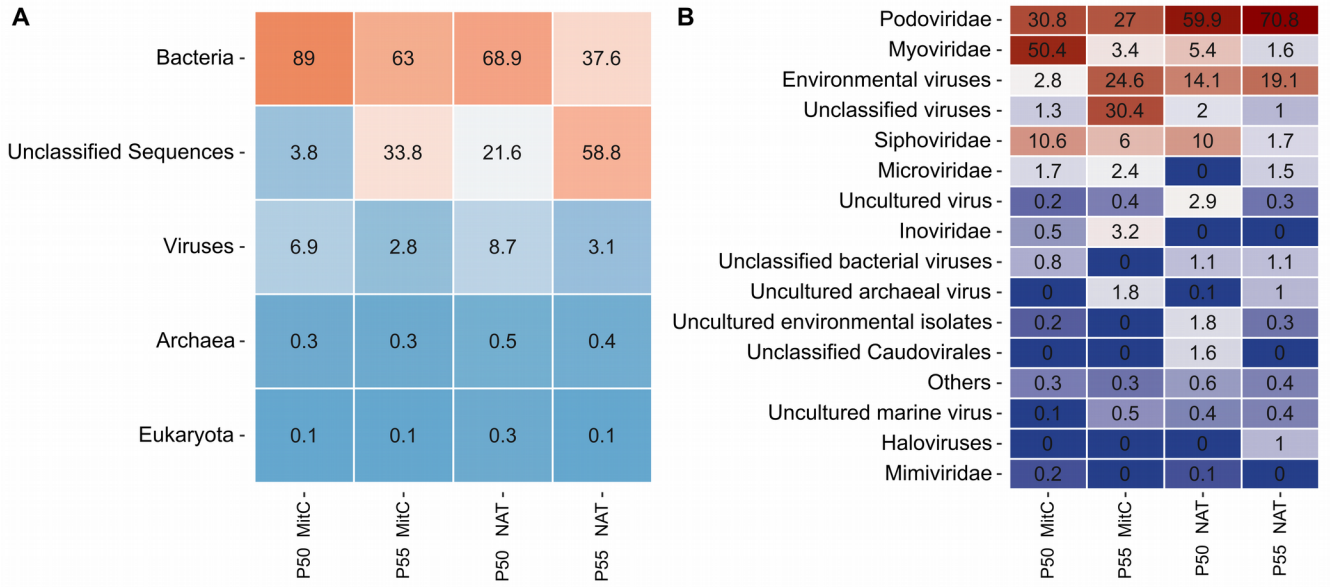


Figure 1. Relative abundances of viral proteins in Porcelana hot spring from two sites and mitomycin C induced and natural samples, classified by LCA algorithm trough local alignment to NCBI nr database. A) Domain level, and **B)** Family level for sequences classified as Virus in A. Sequences were normalized by protein length and library size. Families with abundances below 0.1% were summarized as others.

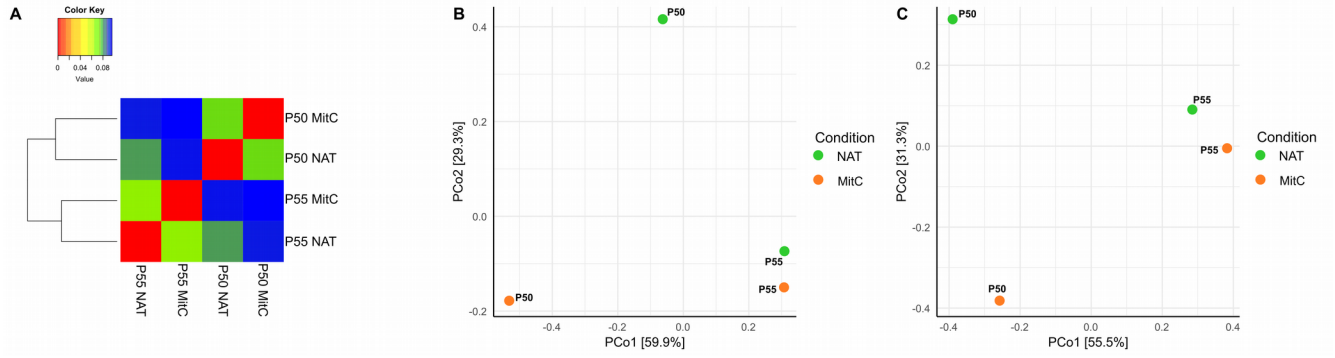


Figure 2. Genetic and compositional dissimilarities of Porcelana hot spring viral communities, measured through *k-mer* frequencies, vPCs and vOTUs based on to MASH and Bray-Curtis distance matrices A)Hierarchical clustering of mitomycin C induced and natural samples from two sites**.** Dendogram was constructed based on to MASH distance matrix and heatmap is colored according to MASH dissimilarity. **B)** Principal coordinate analyses of mitomycin C induction and natural samples from two sites based on vPCs Bray-Curtis distance matrix. **C)** Principal coordinate analyses of mitomycin C induced and natural samples from two sites based on vOTUs Bray-Curtis distance matrix. For both PCoAs no initial data transformation has been applied. The relative contribution (eigenvalue) of each axis to the total inertia in the data is indicated in percent at the axis titles.

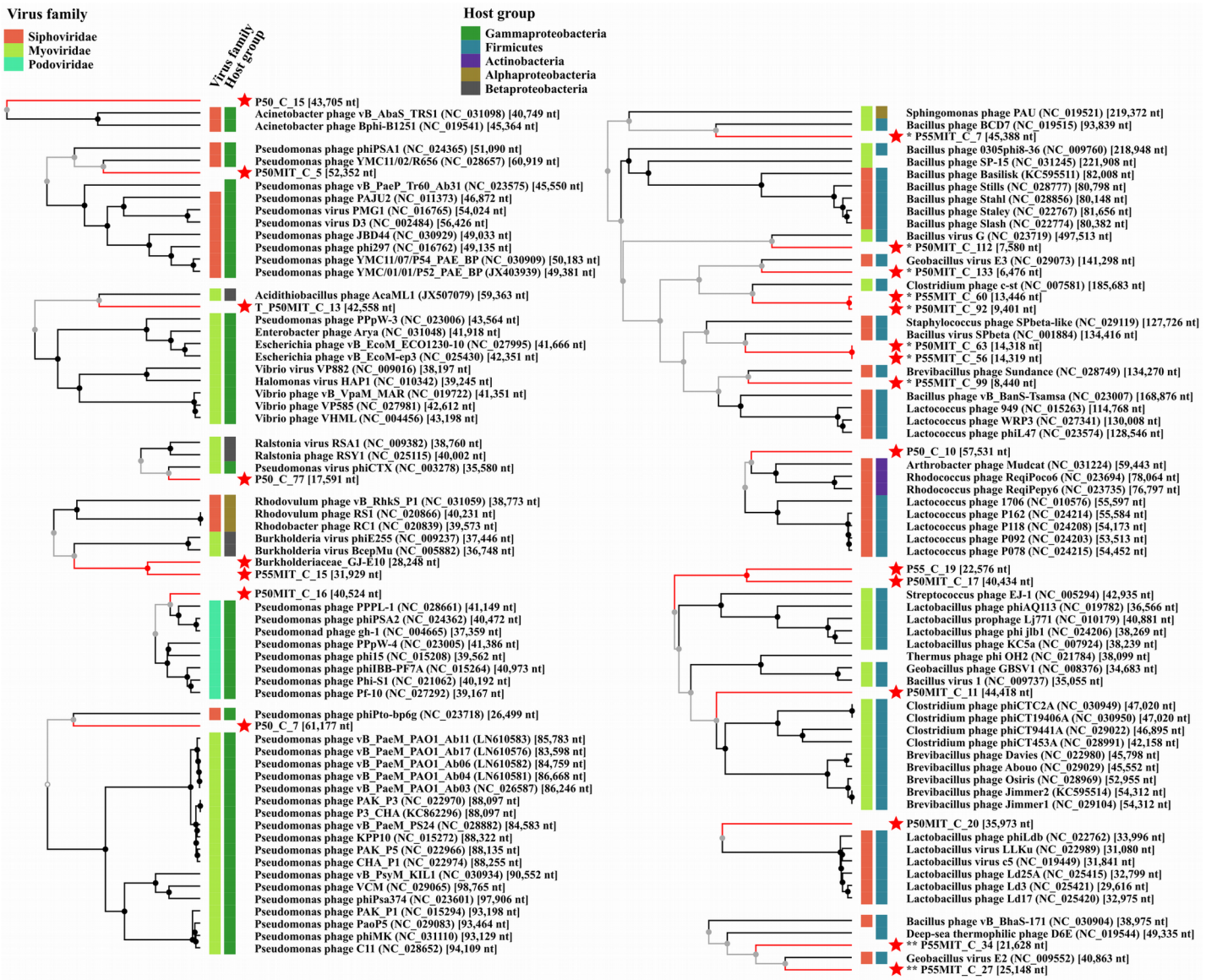


Figure 3. Viral proteomic tree of lysogenic related vOTUs, obtained from PHASTER or differential abundance analyses. Sequences of interest are marked by a red star in the tree, vOTUs obtained by the differential abundance analyses are marked by an asterisk symbol (*) and vOTUs obtained by both analyses are marked by double asterisk symbol (**).

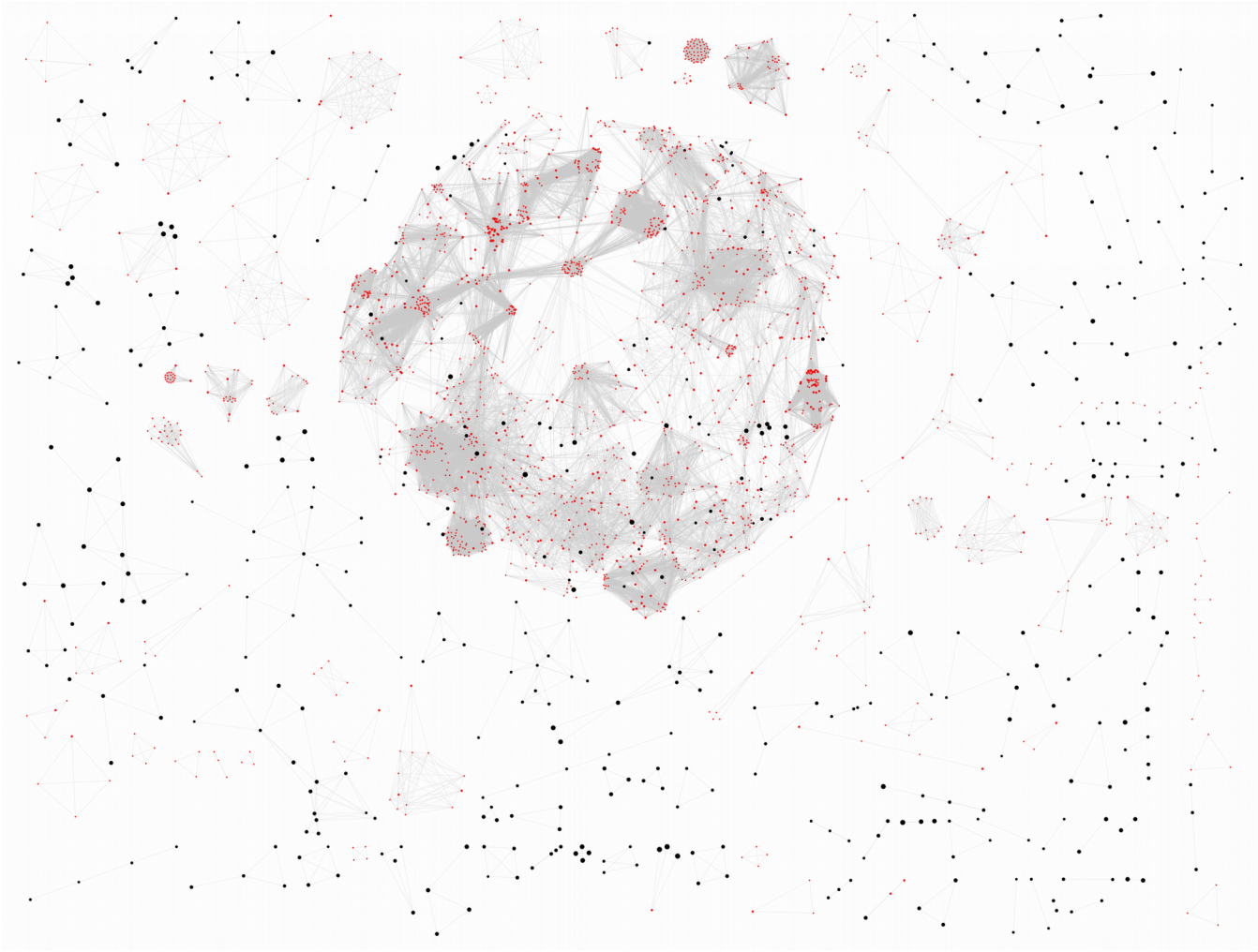


Figure 5. Protein-sharing network for 455 vOTUs and 2243 reference genomes (RefSeq ICTV). Each node represents a vOTU or reference genome, black color represent vOTUs and red color represent reference genomes, node size represent the logarithmic transformation of the vOTUs relative abundance across all samples (four samples). Edges between nodes indicate a statistically significant relationship between the protein profiles of their viral genomes. Modules within the network are composed of groups of similar sequences using the ClusterOne algorithm.

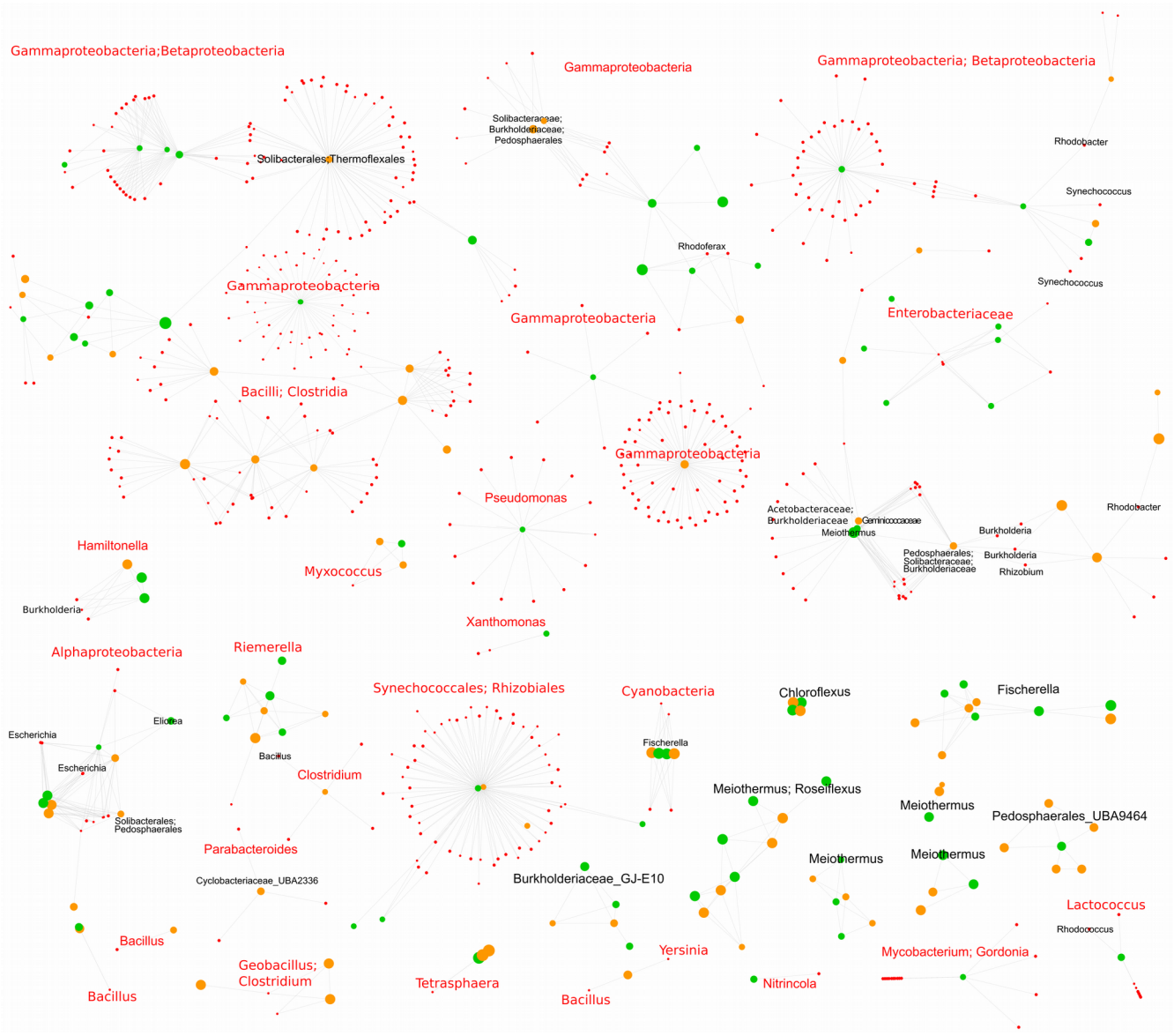


Figure 6. Protein-sharing sub-network for vOTUs with host information. Each node represents a vOTU or reference genome. Orange color represent vOTUs obtained from mitomycin C induced communities, while green color represent vOTUs obtained from natural communities. Red color represent reference genomes. Host classification represented in red letters were obtained from reference viral genomes, while host names in black letters were obtained by CRISPR spacers or *Hexa-mers* frequencies comparison to Porcelana MAGs. Node size represent the logarithmic transformation of the vOTUs relative abundance across all samples (four samples). Edges between nodes indicate a statistically significant relationship between the protein profiles of their viral genomes. Modules within the network are composed of groups of similar sequences using the ClusterOne algorithm.

TABLES

Table 1. Temperate virus search analyses of Metagenome Assembled Genomes (MAGs) from Porcelana hot springs.
Table show taxonomic classification, GC content and length of each MAG. Temperate viruses region were classified as intact or incomplete based on PHASTER score, relevant characteristic of temperate viruses regions such as length, GC content, presence of tRNAs and attachment sites appear in the table.

MAG	Phylum	Order	Family	*Genera*	Length (Mb)	%GC	Viral regions	Completeness	Length (Kb)	%GC	tRNA	Attachment site	Total proteins	Viral proteins	Relevant proteins
1	Acidobacteria	Solibacterales	Solibacteraceae	-	4.17	64.37	2	Incomplete	14.9	65.7	0	yes	20	10	Integrase
									9.9	68.8	0	no	10	9	Tail; Virion
2	Armatimonadetes	Fimbriimonadales	GBS-DC	-	3.82	53.88	1	Incomplete	13.4	63.25	0	no	14	10	Virion
3	Armatimonadetes	Fimbriimonadales	GBS-DC	*GBS-DC*	2.46	61.75	1	Incomplete	14.3	59.97	0	no	16	9	NA
4	Armatimonadetes	Fimbriimonadales	GBS-DC	*GBS-DC*	2.66	60.93	1	Incomplete	7.1	54.13	0	no	10	7	Portal; Head; Capsid
5	Bacteroidetes	Chitinophagales	Saprospiraceae	*UBA10441*	3.79	54.72	0	-	-	-	-	-	-	-	-
6	Bacteroidetes	Chlorobiales	Chloroherpetonaceae	-	3.04	47.75	0	-	-	-	-	-	-	-	-
7	Bacteroidetes	Cytophagales	Cyclobacteriaceae	*UBA2336*	3.48	47.33	1	Incomplete	7.3	43.47	0	no	10	6	NA
8	Bacteroidetes	Kapabacteriales	NICIL-2	*NICIL-2*	2.62	56.12	0	-	-	-	-	-	-	-	-
9	Bacteroidetes	Cytophagales	Cyclobacteriaceae	*UBA2336*	3.15	46.37	1	Incomplete	8.5	46.37	0	no	8	7	Lysin; cI-like repressor
10	Bacteroidetes	Chitinophagales	Saprospiraceae	*UBA10441*	3.65	54.79	0	-	-	-	-	-	-	-	-
11	Chloroflexi	Anaerolineales	SBR1031	A4b	4.46	54.19	1	Incomplete	6.3	54.61	0	no	8	6	Head
12	Chloroflexi	Thermoflexales	-	-	4.35	63.79	1	Incomplete	9.8	63.22	0	no	7	6	Tail
13	Chloroflexi	Thermoflexales	-	-	4.30	63.79	2	Incomplete	9.8	63.22	0	no	7	6	Tail
									11.2	65.53	0	no	9	6	NA
14	Chloroflexi	Thermoflexales	-	-	5.30	64.68	4	Incomplete	9.6	65.3	0	no	7	6	Tail
									10	62.29	0	no	11	8	NA
									9	68.8	0	no	7	6	Tail
									10.1	67.06	0	no	14	8	Transposase
15	Chloroflexi	Chloroflexales	Roseiflexaceae	*Roseiflexus*	4.88	59.37	1	Incomplete	9.6	58.14	1	no	9	6	NA
16	Chloroflexi	Chloroflexales	Roseiflexaceae	*Roseiflexus*	4.86	59.44	2	Incomplete	9.6	58.15	1	no	9	6	NA
									6.1	55.39	1	no	12	6	Capsid; Terminase
17	Chloroflexi	Chloroflexales	Roseiflexaceae	*Roseiflexus*	4.97	58.15	3	Incomplete	9.6	58.15	1	no	9	6	NA
									10.8	43.45	0	no	13	7	Lysin; Tail
									14.7	43.08	0	no	10	7	Tail; Capsid; Portal; Terminase

18	Chloroflexi	Chloroflexales	Chloroflexaceae	*Chloroflexus*	4.82	55.98	1	Incomplete	5.5	54.25	0	no	6	6	NA
19	Chloroflexi	Chloroflexales	Chloroflexaceae	*Chloroflexus*	4.83	55.85	1	Incomplete	5.5	54.25	0	no	6	6	NA
20	Chloroflexi	Chloroflexales	Chloroflexaceae	*Chloroflexus*	4.98	55.85	4	Incomplete	13.1	47.39	0	no	13	8	Terminase; portal; head; capsid
									15.2	52.4	0	yes	19	12	Integrase
									5.5	54.25	0	no	6	6	NA
									6.2	54.15	0	no	10	7	Capsid; Head; Portal
21	Cyanobacteria	Pseudophormidiales	Pseudophormidiaceae	-	6.02	53.53	1	Incomplete	6	55.5	0	no	6	6	NA
22	Cyanobacteria	Cyanobacteriales	Nostocaceae	*Fischerella*	5.49	40.97	1	Incomplete	8.6	39.58	0	no	8	6	NA
23	Cyanobacteria	Cyanobacteriales	Nostocaceae	*Fischerella*	5.26	41.1	4	Incomplete	7.6	40.78	0	no	9	6	NA
									11.1	38.62	0	yes	8	6	Integrase-Resolvase; Transposase
									8.3	39.17	0	no	8	6	NA
									5.5	47.12	0	no	6	6	NA
24	Cyanobacteria	Cyanobacteriales	Nostocaceae	*Fischerella*	5.40	41.11	3	Incomplete	25.8	40.14	0	yes	8	6	Protease; Integrase; Transposase
									7.6	40.79	0	no	9	6	NA
									8.6	39.58	0	no	9	6	NA
25	Deinococcus-Thermus	Deinococcales	Thermaceae	*Meiothermus*	3.89	59.25	3	Incomplete	14.7	43.12	0	no	10	7	Terminase; Portal ; Capsid; Tail
									10.8	43.51	0	no	12	7	Tail; Lysin
									21.4	62.98	0	yes	7	6	Integrase
26	Planctomycetes	Isosphaerales	Isosphaeraceae	-	5.69	64.35	1	Incomplete	11.5	66.37	0	no	14	10	Head; Capsid; Tail
25	Planctomycetes	Phycisphaerales	UBA1161	-	3.49	52.45	1	Incomplete	8.3	50.87	0	no	8	6	DprA; Tail fiber
28	Planctomycetes	Phycisphaerales	UBA1161	-	3.30	52.45	1	Incomplete	8.5	50.06	0	no	8	6	Tail
29	Planctomycetes	Phycisphaerales	UBA1161	-	3.44	52.76	3	Incomplete	8.5	50.06	0	no	8	6	Tail
								Incomplete	8.3	50.87	0	no	8	6	DprA; Tail fiber
								Incomplete	7.6	72.17	0	no	9	7	Tail fiber
30	Proteobacteria	Betaproteobacteriales	Burkholderiaceae	*GJ-E10*	2.97	68.57	1	Intact	28.2	68.61	0	yes	43	26	Virion; Tail; Capsid; Portal; Transposase; Integrase
31	Proteobacteria	Acetobacterales	Acetobacteraceae	*Elioraea*	4.01	71.41	1	Incomplete	18.5	71.71	0	no	24	18	Capsid; Tail
32	Proteobacteria	Acetobacterales	Acetobacteraceae	*Elioraea*	4.02	72.59	3	Incomplete	7.7	74.09	0	no	11	6	Lysin; Terminase;

															cI-like repressor
									4.5	68.68	0	no	9	6	Transposase
									6.7	69.44	0	no	9	7	NA
33	Proteobacteria	Geminicoccales	Geminicoccaceae	-	3.87	72.68	2	Incomplete	9.9	71.24	0	no	11	7	Head; Capsid; Tail
									12.2	69.68	0	no	17	10	NA
34	Verrucomicrobia	Pedosphaerales	UBA9464	-	4.20	67.14	1	Incomplete	6	66.07	0	no	9	6	Transposase

Table 2. Results of two way Brunner-Dette-Munk test of Porcelana vPCs abundances from two sites and mitomycin C induced and natural samples.

Factor	df1	df2	F*	P (F > F*)
Condition	1	37083.31	28.31	1.04×10^{-07}
Site	1	37083.31	801.66	1.70×10^{-174}
Condition:Site	1	37083.31	214.18	2.30×10^{-48}

Table 3: Results of Wilcoxon pairwise comparisons rank sum test of Porcelana vPCs abundances from two sites and mitomycin C induced and natural samples.

Groups	P-value
P50MitC:P50NAT	3.22×10^{-51}
P55NAT:P50NAT	5.36×10^{-185}
P55NAT:P50MitC	6.14×10^{-61}
P55MitC:P50NAT	7.18×10^{-115}
P55MitC:P50MitC	2.64×10^{-25}
P55MitC:P55NAT	1.89×10^{-12}

Table 4: Results of two way Brunner-Dette-Munk test of Porcelana vOTUs abundances from two sites and mitomycin C induced and natural samples.

Factor	df1	df2	F*	P (F > F*)
Condition	1	2948.89	0.45	0.50
Site	1	2948.89	21.28	4.14×10^{-06}
Condition:Site	1	2948.89	27.61	1.59×10^{-07}

Table 5: Results of Wilcoxon pairwise comparisons rank sum test of Porcelana vOTUs abundances from two sites and mitomycin C induced and natural samples.

Groups	P-value
P50NAT:P50MitC	0.0016
P50NAT:P55MitC	0.0454
P55NAT:P50MitC	0.0016
P55NAT:P55MitC	0.0454

Table 6: Summary results of the viral proteomic tree analysis of the most abundant (≥1%) vOTUs in Porcelana Hot springs. Table show relative abundance of vOTUs in each sample and percent of relative abundance in the total vOTUs set. Taxonomic classification of the Porcelana MAG is provided for vOTUs that have a CRISPR spacer hit. Closest reference viral genome was obtained from SG score of the proteomic tree analysis.

vOTU	P50 (counts)	P50MIT (counts)	P55 (counts)	P55MIT (counts)	Total abundance (%)	CRISPR	Closest reference viral genome	Reference Lifestyle	Reference host phylum	Reference ICTV classification
P55_C_19	16	2973	3299	527	3.3	--	Streptococcus phage phiD12	Lysogenic	Firmicutes	Caudovirales
P50MIT_C_6	52	6086	0	8	2.9	--	Tetrasphaera phage TJE1	Lytic	Actinobacteria	Caudovirales
P55MIT_C_21	61	5932	0	26	2.9	--	Tetrasphaera phage TJE1	Lytic	Actinobacteria	Caudovirales
P50_C_11	51	5606	0	7	2.7	--	Tetrasphaera phage TJE1	Lytic	Actinobacteria	Caudovirales
P55MIT_C_169	298	468	2445	1439	2.2	--	Thermoanaerobacterium phage THSA-485A	Lytic	Firmicutes	Siphoviridae
P50_C_225	11	64	1728	2755	2.2	--	Rhodococcus phage E3	unknown	Actinobacteria	Myoviridae
P50_C_52	2508	450	0	0	1.4	--	Acidithiobacillus phage	Lytic	Firmicutes	Myoviridae

							AcaML1			
P55_C_104	1827	924	106	96	1.4	--	unknown	unknown	unknown	unknown
P50_C_26	1061	575	888	244	1.3	Fischerella	Phormidium virus WMP3	Lytic	Cyanobacteria	Podoviridae
P55MIT_C_12	868	500	1076	307	1.3	Fischerella	Phormidium virus WMP3	Lytic	Cyanobacteria	Podoviridae
P55MIT_C_163	1634	897	93	105	1.3	--	unknown	unknown	unknown	unknown
P55_C_10	820	500	1058	280	1.3	Fischerella	Phormidium virus WMP3	Lytic	Cyanobacteria	Podoviridae
P55_C_79	2386	181	35	52	1.3	--	Tsukamurella phage TPA4	Lytic	Actinobacteria	Siphoviridae
P50MIT_C_23	927	541	915	241	1.3	Fischerella	Phormidium virus WMP3	Lytic	Cyanobacteria	Podoviridae
P55MIT_C_71	168	264	1459	723	1.3	--	Thermoanaerobacterium phage THSA-485A	Lytic	Firmicutes	Siphoviridae
P50MIT_C_30	0	2488	0	0	1.2	--	Pseudomonas virus DMS3	Lysogenic	Proteobacteria	Siphoviridae
P55_C_43	154	230	1358	692	1.2	--	Thermoanaerobacterium phage THSA-485A	Lytic	Firmicutes	Siphoviridae
P50MIT_C_69	146	240	1340	690	1.2	--	Thermoanaerobacterium phage THSA-485A	Lytic	Firmicutes	Siphoviridae
P50_C_42	1062	902	49	124	1.0	Fischerella	Methanobacterium phage psiM2	Lytic	Methanobacteria	Siphoviridae
P55MIT_C_22	1062	899	47	124	1.0	Fischerella	Methanobacterium phage psiM2	Lytic	Methanobacteria	Siphoviridae
P50_C_313	87	47	1098	838	1.0	Meiothermus	Streptomyces phage mu1/6	Lysogenic	Actinobacteria	Siphoviridae
P55_C_100	1856	91	27	35	1.0	--	Tsukamurella phage TPA2	Lytic	Actinobacteria	Siphoviridae

Supplementary Material

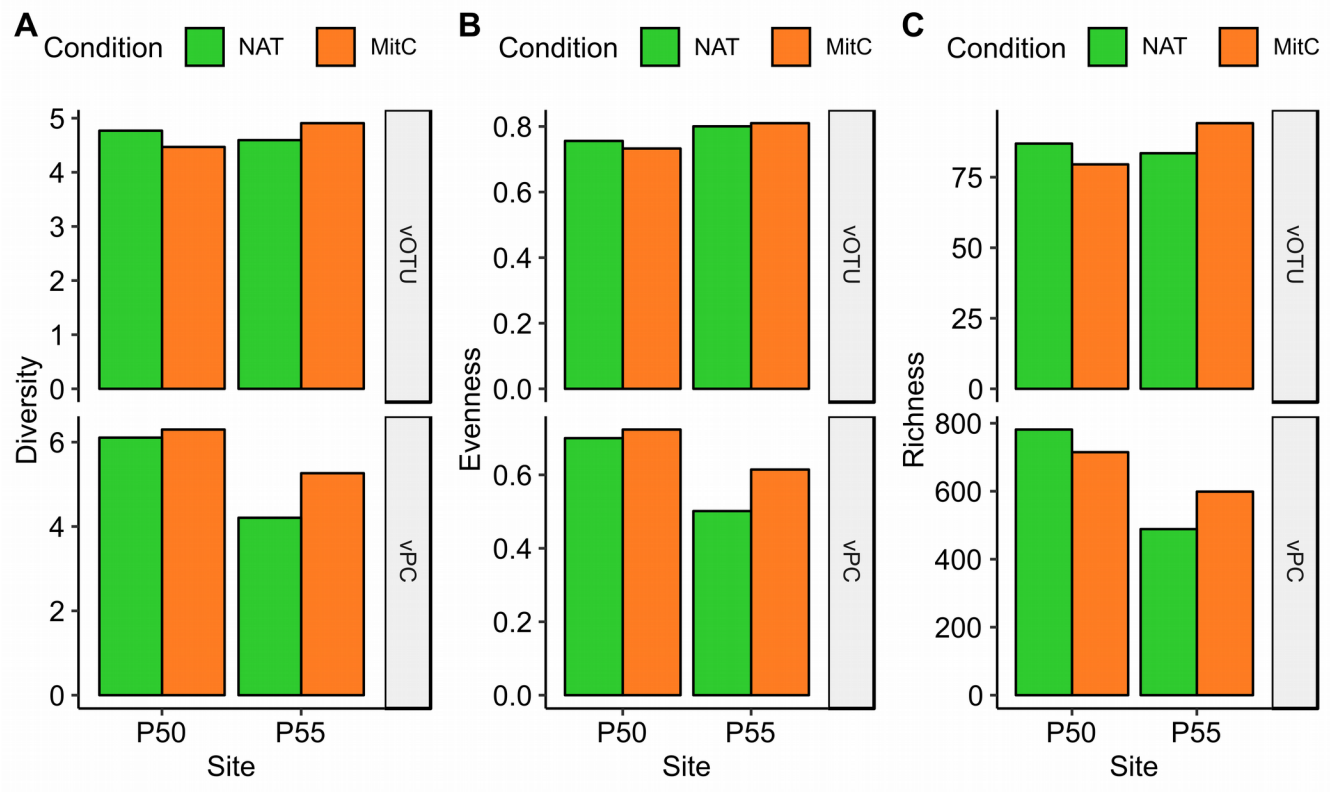
Killing the winner and piggybacking the cheater: lytic and lysogenic viral communities in hot springs phototrophic mats.

Sergio Guajardo-Leiva, Oscar Salgado, and Beatriz Díez.

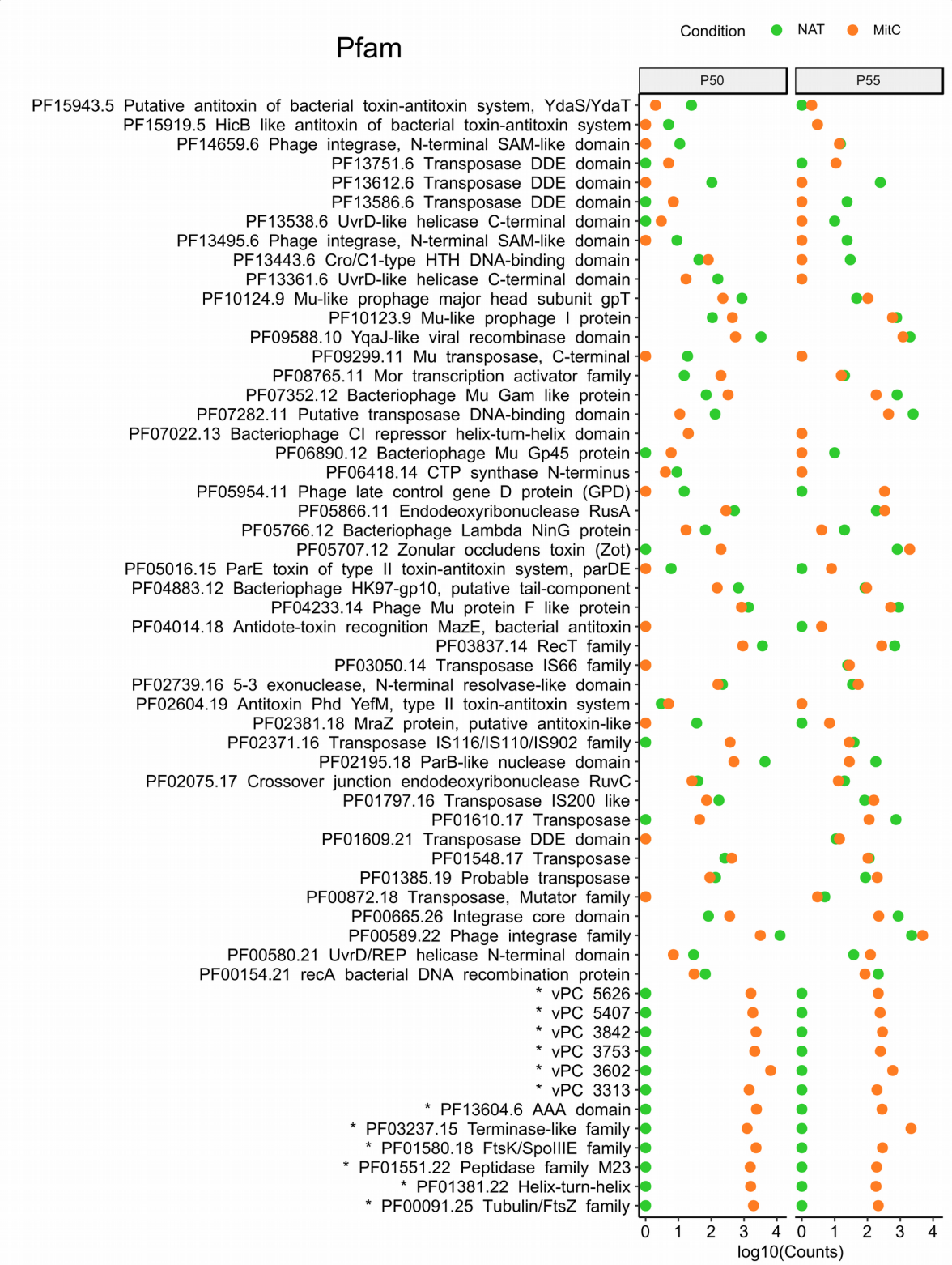
*** Correspondence:** Beatriz Díez: bdiez@bio.puc.cl

1 Supplementary Figures and Tables

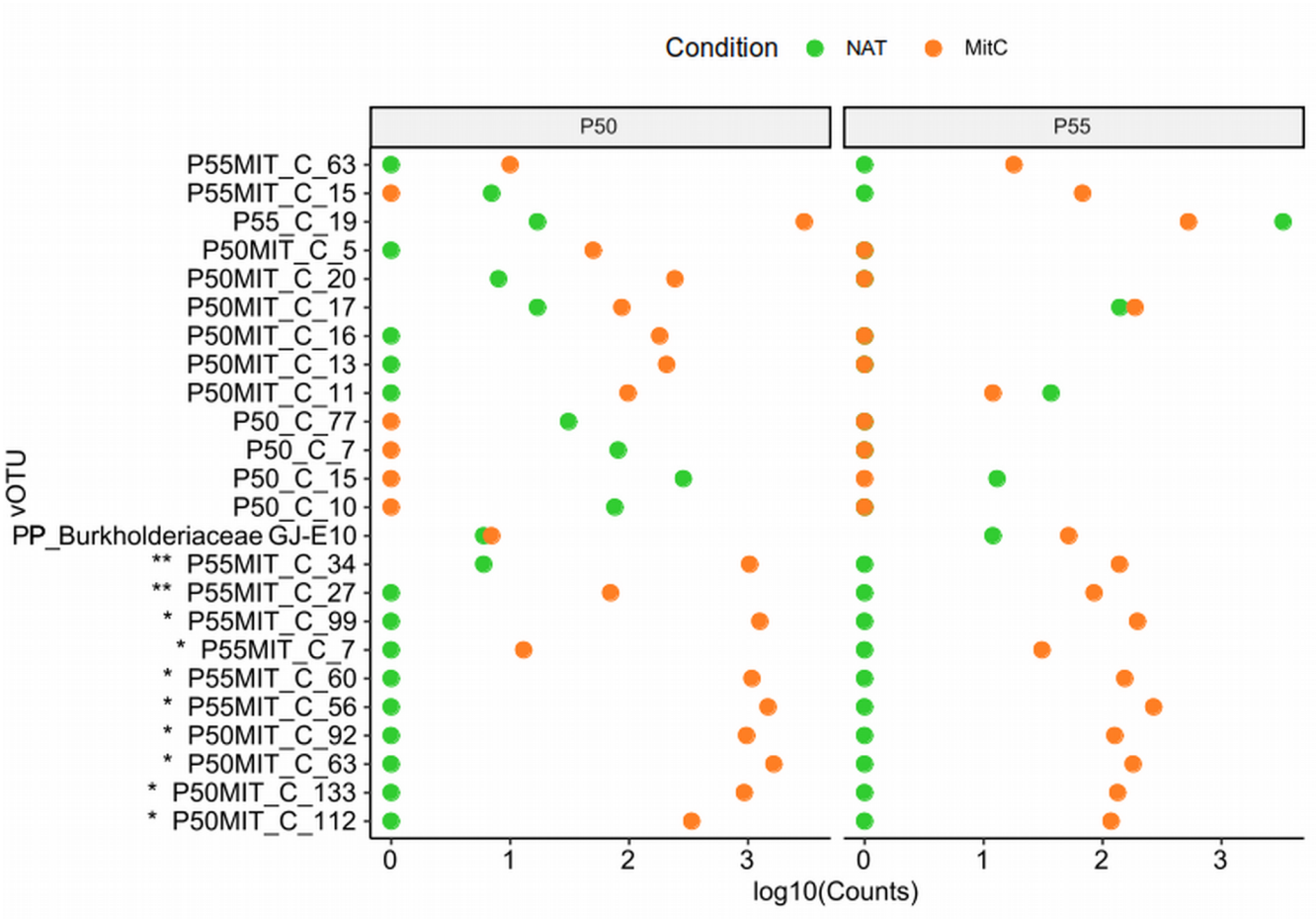
1.1 Supplementary Figures



Supplementary Figure S1.Alpha diversity of Porcelana hot spring natural and mitomycin C induced communities at two sites. vPCs and vOTUs normalized counts for each sample were used to calculate A) Shannon's diversity, B) Pielou's evenness and C) Species Richness.



Supplementary Figure S2. Relative abundance of lysogenic related vPCs, obtained from Pfam annotation or differential abundance analyses. Viral PCs obtained by the differential abundance analyses are marked by an asterisk symbol (*).



Supplementary Figure S3. Relative abundances of lysogenic vOTUs, obtained from PHASTER annotation or differential abundance analyses. Viral OTUs obtained by the differential abundance analyses are marked by an asterisk symbol (*) and vOTUs obtained by both techniques are marked by double asterisk symbol (**).

1.2 Supplementary Tables.

Supplementary Table S1. Summary information about sequencing depth, quality filtering, read mapping and assembly of hot springs viral metagenomes.

Sample	Raw sequences (M)	Sequences (M) after quality filter	Bases (M) after quality filter	Number of viral contigs ≥ 500pb	Reads in contigs ≥ 500pb (M)	Number of predicted proteins	nr aligned sequences (M)	vPCs recruited sequences (M)	vOTUs recruited sequences (M)
P50_NAT	4.13	3.69	859.60	11961	1.91	27728	0.83	1.37	1.62
P55_NAT	3.01	2.77	655.01	6197	1.72	13087	0.87	1.46	0.68
P50_MitC	8.67	7.90	1811.56	18303	2.81	34233	1.08	2.09	3.09
P55_MitC	6.91	6.32	1456.28	14306	3.49	27677	1.60	2.98	2.28

Supplementary Table S2. Relative abundances of vPCs with Pfam annotation.
Counts from protein clusters with the same Pfam accesion were added in each sample. Only functions with relative abundances above 0.1% of the total vPCs counts are showed.

Pfam	Accession	P50NAT	P55NAT	P50MitC	P55MitC	Total
Phage integrase family	PF00589.22	12469	2217	3140	4801	22627
Terminase RNaseH-like domain	PF17289.2	11431	3577	3792	1808	20608
DNA polymerase family A	PF00476.20	5189	2139	3210	6253	16791
Domain of unknown function (DUF4774)	PF15999.5	690	417	1154	13290	15551
Peptidase family M23	PF01551.22	5703	3044	2933	1638	13318
Replication initiation factor	PF02486.19	12055	146	452	226	12879
Bacteriophage head to tail connecting protein	PF12236.8	5668	2586	1388	1131	10773
dUTPase	PF00692.19	4341	1516	1175	2663	9695
N-acetylmuramoyl-L-alanine amidase	PF01510.25	4157	274	1622	3371	9424
Terminase-like family	PF03237.15	858	1408	1604	5343	9213
Poxvirus A32 protein	PF04665.12	431	2477	0	6178	9086
NlpC/P60 family	PF00877.19	49	0	8723	9	8781
Tail tubular protein	PF17212.3	5281	1690	1171	475	8617
Phage portal protein	PF04860.12	5101	257	1800	310	7468
DnaB-like helicase C terminal domain	PF03796.15	4765	97	2347	135	7344
YqaJ-like viral recombinase domain	PF09588.10	3293	1979	547	1212	7031
Phage capsid family	PF05065.13	5042	177	1535	275	7029
Protein of unknown function (DUF1071)	PF06378.11	3792	1760	989	442	6983
N-acetylmuramoyl-L-alanine amidase	PF01520.18	4039	1362	1034	463	6898
Phage terminase large subunit (GpA)	PF05876.12	4455	1064	972	389	6880
Phage-related minor tail protein	PF10145.9	460	1131	763	4316	6670
Thymidylate kinase	PF02223.17	3377	1184	995	415	5971
FtsK/SpoIIIE family	PF01580.18	235	83	2505	3088	5911
Ribonucleotide reductase, barrel domain	PF02867.15	537	207	502	4415	5661
Bacteriophage T4-like capsid assembly protein (Gp20)	PF07230.11	192	149	305	4930	5576
RecT family	PF03837.14	3633	678	917	270	5498
Helix-turn-helix domain	PF12728.7	3923	210	993	156	5282
Glycosyl transferases group 1	PF00534.20	180	170	274	4496	5120
ParB-like nuclease domain	PF02195.18	4339	179	488	27	5033
Major capsid protein Gp23	PF07068.11	148	136	273	4447	5004
Prohead core protein serine protease	PF03420.13	171	120	308	4362	4961

Caudovirus prohead serine protease	PF04586.17	3412	71	1263	213	4959
Protein of unknown function (DUF3987)	PF13148.6	4131	64	439	128	4762
VRR-NUC domain	PF08774.11	2876	161	1437	189	4663
AAA domain	PF13481.6	3042	122	1239	212	4615
Calcineurin-like phosphoesterase	PF00149.28	827	912	1810	1049	4598
Prophage endopeptidase tail	PF06605.11	89	100	1168	3147	4504
Bacterial regulatory protein, Fis family	PF02954.19	3106	55	1104	126	4391
Phage tail protein	PF05709.11	146	120	241	3838	4345
Phage portal protein, lambda family	PF05136.13	3938	0	320	5	4263
Concanavalin A-like lectin/glucanases superfamily	PF13385.6	225	158	228	3429	4040
PD-(D/E)XK nuclease superfamily	PF12705.7	575	628	1587	1128	3918
VirE N-terminal domain	PF08800.10	118	1997	436	1092	3643
Phage Mu protein F like protein	PF04233.14	1364	895	836	510	3605
Putative phage serine protease XkdF	PF14550.6	583	1674	345	940	3542
Toprim-like	PF13155.6	127	91	192	2876	3286
Uncharacterized conserved protein (DUF2190)	PF09956.9	2796	40	285	100	3221
Bacteriophage holin family	PF05105.12	2235	33	845	87	3200
AAA domain (dynein-related subfamily)	PF07728.14	2582	157	270	100	3109
Putative transposase DNA-binding domain	PF07282.11	131	2468	10	439	3048
Meiotically up-regulated gene 113	PF13455.6	176	89	160	2602	3027
Helix-turn-helix	PF01381.22	576	180	1854	375	2985
Zonular occludens toxin (Zot)	PF05707.12	0	811	196	1939	2946
Trypsin-like peptidase domain	PF13365.6	2573	33	248	65	2919
Kelch motif	PF13964.6	74	936	153	1613	2776
Protein of unknown function (DUF935)	PF06074.12	219	1040	576	933	2768
AAA domain	PF13604.6	0	6	2392	278	2676
Glycosyl hydrolase 108	PF05838.12	493	661	1425	39	2618
VWA-like domain (DUF2201)	PF09967.9	2380	0	175	0	2555
Helix-turn-helix domain	PF13560.6	1961	19	461	49	2490
Phage terminase large subunit	PF04466.13	794	306	727	596	2423
Reverse transcriptase (RNA-dependent DNA polymerase)	PF00078.27	2083	0	315	0	2398
SNF2 family N-terminal domain	PF00176.23	258	393	1353	391	2395
CHC2 zinc finger	PF01807.20	1846	97	289	57	2289
Tubulin/FtsZ family, GTPase domain	PF00091.25	0	0	1936	212	2148
Large polyvalent protein associated domain 38	PF18857.1	127	798	116	1106	2147
Helicase conserved C-terminal domain	PF00271.31	92	1117	0	766	1975
Baseplate J-like protein	PF04865.14	62	1098	406	380	1946
Putative amidoligase enzyme	PF12224.8	257	579	262	847	1945
Mu-like prophage I protein	PF10123.9	107	775	440	580	1902
AAA domain	PF13401.6	240	827	509	221	1797
DNA methylase	PF01555.18	1244	133	93	263	1733
Mitochondrial genome maintenance MGM101	PF06420.12	291	562	101	763	1717
Helix-turn-helix domain	PF13730.6	311	535	91	778	1715
P22 coat protein - gene protein 5	PF11651.8	1139	112	373	37	1661
Protein of unknown function (DUF2800)	PF10926.8	270	83	1255	17	1625
Phage virion morphogenesis family	PF05069.13	99	61	1306	77	1543
Integrase core domain	PF00665.26	81	864	362	220	1527
N-terminal phage replisome organiser (Phage rep org N)	PF09681.10	93	0	1426	0	1519
Phage portal protein, SPP1 Gp6-like	PF05133.14	880	139	320	162	1501
Protein of unknown function (DUF2815)	PF10991.8	103	46	1255	4	1408

Chaperonin 10 Kd subunit	PF00166.21	1112	46	181	56	1395
Bacteriophage Mu Gam like protein	PF07352.12	69	798	324	183	1374
Helix-turn-helix domain	PF12844.7	96	784	292	197	1369
Endodeoxyribonuclease RusA	PF05866.11	508	186	280	336	1310
C-5 cytosine-specific DNA methylase	PF00145.17	1034	116	95	30	1275
Mu-like prophage major head subunit gpT	PF10124.9	861	46	225	102	1234
HNH endonuclease	PF01844.23	185	564	290	183	1222
DNA primase catalytic core, N-terminal domain	PF08275.11	0	0	1034	130	1164
Protein of unknown function (DUF1320)	PF07030.12	167	417	172	405	1161
DNA polymerase type B, organellar and viral	PF03175.13	89	370	36	596	1091
Protein of unknown function (DUF1018)	PF06252.12	45	602	250	160	1057
D12 class N6 adenine-specific DNA methyltransferase	PF02086.15	441	227	105	281	1054
Capsid protein (F protein)	PF02305.17	0	107	464	482	1053
Protein of unknown function (DUF3168)	PF11367.8	749	42	125	125	1041
D-alanyl-D-alanine carboxypeptidase	PF13539.6	66	4	946	10	1026
Bacteriophage HK97-gp10, putative tail-component	PF04883.12	673	83	151	93	1000
3D domain	PF06725.11	0	0	889	107	996
ASCH domain	PF04266.14	880	30	71	5	986
Bacterial dnaA protein helix-turn-helix	PF08299.11	255	267	447	17	986
Putative metallopeptidase domain	PF13203.6	870	0	101	3	974
Phage major capsid protein E	PF03864.15	17	181	391	374	963
Phage Tail Collar Domain	PF07484.12	351	131	372	101	955
Polysaccharide deacetylase	PF01522.21	729	8	204	0	941
Transposase	PF01548.17	259	112	420	103	894
Transposase	PF01610.17	0	739	43	111	893
Family of unknown function (DUF5309)	PF17236.2	558	0	169	134	861
Bacterial regulatory protein, arsR family	PF01022.20	25	518	195	103	841
Putative bacterial sensory transduction regulator	PF10722.9	746	16	63	14	839
Replication initiation and membrane attachment	PF07261.11	673	14	132	15	834
DNA N-6-adenine-methyltransferase (Dam)	PF05869.11	50	275	5	501	831
Type III restriction enzyme, res subunit	PF04851.15	188	193	16	413	810
Calcineurin-like phosphoesterase superfamily domain	PF12850.7	132	182	26	455	795
Phage lysozyme	PF00959.19	80	56	504	109	749
ERF superfamily	PF04404.12	483	27	215	13	738
Phage P22-like portal protein	PF16510.5	695	11	21	3	730
Terminase small subunit	PF03592.16	176	153	180	213	722
Rad52/22 family double-strand break repair protein	PF04098.15	81	191	67	369	708
Phage head-tail joining protein	PF05521.11	499	43	125	34	701
CHAP domain	PF05257.16	622	7	60	7	696
Anaerobic ribonucleoside-triphosphate reductase	PF13597.6	28	366	176	107	677
P22 tail accessory factor	PF11650.8	342	64	185	62	653
Peptidase S24-like	PF00717.23	54	93	418	78	643
Protein of unknwon function (DUF3310)	PF11753.8	18	20	562	0	600
Putative phage tail protein	PF13550.6	71	130	39	344	584
Uracil DNA glycosylase superfamily	PF03167.19	221	100	143	117	581
Clp protease	PF00574.23	220	248	36	63	567
N-acetylmuramidase	PF11860.8	148	88	161	152	549
Peptidase C26	PF07722.13	123	68	273	83	547
Glutamine amidotransferase class-I	PF00117.28	64	60	228	189	541
Primase C terminal 1 (PriCT-1)	PF08708.11	525	0	0	0	525

Methyltransferase domain	PF08241.12	462	12	38	13	525
Glycosyltransferase family 29 (sialyltransferase)	PF00777.18	475	0	40	0	515
Probable transposase	PF01385.19	134	86	90	198	508
Ankyrin repeats (3 copies)	PF12796.7	98	64	0	343	505

1305
1306
1307

CHAPTER 3

Ecological drivers modulate biogeography in thermophilic viral communities.

Ecological drivers modulate biogeography in thermophilic viral communities

Sergio Guajardo-Leiva[1], Tomás Alarcón-Schumacher[1-2], Oscar Salgado[1], Juris A. Grasis[3], Carlos Pedrós-Alió[4], Forest Rohwer[3] and Beatriz Díez[1,5]

[1]Department of Molecular Genetics and Microbiology, Pontificia Universidad Católica de Chile, Santiago, Chile.

[2]Max Planck Institute for Marine Microbiology (MPG), Bremen, Germany.

[3]Department of Biology, San Diego State University, San Diego, CA, USA.

[4]Programa de Biología de Sistemas, Centro Nacional de Biotecnología (CSIC), Madrid, España.

[5]Center for Climate and Resilience Research (CR)2, Chile.

* Correspondence:
Beatriz Díez
bdiez@bio.puc.cl

Hot springs, Biogeography, Viruses, CRISPR , Viral Metagenomics, Viral OTUs, Viral Protein Clusters.

ABSTRACT

Viruses have proven to be ubiquitous in all environments and hot springs are not the exception even though its extreme conditions. Despite the interest to understand how viruses deal with high temperatures, there are no studies addressing environmental determinants over thermophilic viral community structure or their biogeographic patterns.

Here we analyzed the 12 hot springs viral metagenomes that have been published to date, to reveal how viral community structure is affected by ecological drivers that define hot springs over the American continent in a latitudinal scale. The extensive analysis of protein, gene and genome sequences using *k-mer* frequencies, viral Protein Clusters (vPCs) and viral Operational Taxonomic Units (vOTUs) showed a biogeographic pattern according to major ecological drivers (here, pH and temperature). The structure of the viral community was affected by both, pH and temperature, while pH had a stronger effect on the distribution and abundance of the vPCs, with temperature being the factor that most affected the

vOTUs. The protein sharing network of the viral communities, showed an unexpected modularity, that suggest a restriction to gene flow between hot springs. A high local viral richness was associated to specific hosts when crossing CRISPR spacers information and viral modules of the network, noticing the existence of specific virus-host pairs that allow the maintenance of this local richness.

These viral metagenomic analyses of hot springs reveal biogeographic patterns at the community level that suggest passive air transport on a local scale, but also probably on a global scale. Locally, viral communities are structured influenced by environmental conditions (pH and temperature) that primarily affect the structure of the host community.

1. INTRODUCTION

Hot springs represent discontinuous habitats with a rich interphase between aquatic and terrestrial environments, which determines their physicochemical properties. Usually they present geochemical and physical gradients distributed along contrasting distances, with scales that can go from the centimeters to hundreds of meters and kilometers (Papke et al., 2003; Sharp et al., 2014). Microbial communities from these high temperature habitats are usually dominated by few types of microorganisms (some phyla) and usually are less diverse than lower temperature fresh water, terrestrial habitats and oceans (Inskeep et al., 2010, 2013). The relative simplicity of hot spring communities, has allowed its use as models to correlate genomic functions with environmental parameters, and for the understanding of environmental determinants over the community structure (Inskeep et al., 2013; Klatt et al., 2013; Power et al., 2018; Sharp et al., 2014).

Likewise and since hot springs are a discontinuous environment that can be considered as hot islands in a cold ocean world, which acts as an effective environmental filter, they have been used to test the Louren Baas Becking's hypothesis that 'everything is everywhere, but the environment select (O'Malley, 2008). Functional characterization of thermophilic bacterial and archaeal communities have shown, that differences in physicochemical gradients produces differences in specific genes and major functional pathways, which in turn follow the division between bacteria and archaea revealing hallmark signatures of metabolic capabilities (Alcamán-Arias et al., 2018; Inskeep et al., 2010). On the same line, temperature (Sharp et al., 2014), pH (Inskeep et al., 2010, 2013; Power et al., 2018) and sulfide, or elemental sulfur (Inskeep et al., 2010, 2013; Menzel et al., 2015) have been proposed as major drivers

of bacterial and archaeal communities structure in these environments. In general, sites with pH values between 5-9 and a temperature ranges between 45-70 ºC are dominated by phototrophic organisms either oxygenic or anoxygenic, depending on sulfide or elemental sulfur concentrations levels (Alcamán-Arias et al., 2018; Inskeep et al., 2013; Menzel et al., 2015). In the same range of pH, but temperatures > 70 ºC and associated to high concentrations of dissolved sulfide or elemental sulfur, the dominant groups are Aquificales and Thermoproteales (Inskeep et al., 2013; Menzel et al., 2015). Finally, Sulfolabales members dominate the community at acidic pH range (2 to 5) and temperature > 70 ºC (Inskeep et al., 2013; Menzel et al., 2015). On the other hand, the acidic sites (pH 2-5) with mild between 45 to 70 ºC have been systematically under sampled (Wilson et al., 2008).

Otherwise, hot springs viral communities, have been poorly characterized and mostly through studies restricted to low pH and high-temperature (Bolduc et al., 2012, 2015; Gudbergsdóttir et al., 2016; Munson-Mcgee et al., 2018) or at high-temperature and neutral pH hot spring (Breitbart et al., 2004; Davison et al., 2016; Pride and Schoenfeld, 2008; Schoenfeld et al., 2008). In these environments, viruses have proven to be ubiquitous, numerous and actives (Bolduc et al., 2012, 2015; Breitbart et al., 2004; Guajardo-Leiva et al., 2018; Gudbergsdóttir et al., 2016; Menzel et al., 2015; Munson-Mcgee et al., 2018; Schoenfeld et al., 2008; Zablocki et al., 2017), were they shape the diversity of their host through co-evolution (Guajardo-Leiva et al., 2018; Sano et al., 2018), and regulate the structure of cellular communities given that they are usually the only predator above 50 ºC (Breitbart et al., 2004; Klatt et al., 2013; Schoenfeld et al., 2008).

Database dependent analyses from these studies, reported that the structure of viral communities at mild and high-temperature with circumneutral pH hot springs were dominated by bacterial viruses from Myoviridae, Siphoviridae and Podoviridae families (Caudovirales order) (Guajardo-Leiva et al., 2018; Pride and Schoenfeld, 2008; Schoenfeld et al., 2008; Zablocki et al., 2017) as well as Thermoproteales viruses from Globuloviridae family (Pride and Schoenfeld, 2008; Schoenfeld et al., 2008). In particular, some of the most abundant Caudovirales retrieved were cyanophages, viruses which infect cyanobacteria (Guajardo-Leiva et al., 2018; Zablocki et al., 2017). Phylogenetic analyses showed those viruses were part of a monophyletic clade that includes fresh water cyanophages that also infects filamentous cyanobacteria (Guajardo-Leiva et al., 2018). Meanwhile at low pH and high temperature hot springs Sulfolobales viruses were the dominant and usually associated to a single viral family such

as Lipothrixviridae, Fuselloviridae, Ampullaviridae, Bicaudaviridae and Rudiviridae (Gudbergsdóttir et al., 2016).

However, the use of approaches that do not relies in databases, such as those carried in Nymph Lake hot spring, Yellowstone National Park (YNP) allowed to find an underlying diversity of archaeal viruses, not associated to known families, reveling 103 new viral groups that infect 8 different species of Sulfolobales (Bolduc et al., 2015). The use of this type of approach, such as viral contig networks, opens new possibilities in the study of viral communities from less explored environments such as hot springs. Despite the significant contribution of all these studies in the field of environmental viruses in hot springs, most of them have only analyzed one or two single hot springs from the same region, wherein none of them has investigated the relationships between the structure of viral communities and the basic physicochemical parameters such as temperature and pH in hot springs from different geographic locations.

Regardless of the environmental conditions of each thermal site, the resident microorganisms can be dispersed by air due to the existing exchange with the steam emanating from these thermal sources (Herbold et al., 2014). In this sense, there would be multiple factors that would explain the distribution of microorganisms in these systems. For example, there is evidence that supports the idea of niche selection at the local scale but also distance decay patterns and endemism, as well as allopatric speciation, have been described in hot springs (Herbold et al., 2014; Menzel et al., 2015; Papke et al., 2003; Power et al., 2018; Whitaker, 2003).

Biogeography studies of viruses in hot springs are much more limited than those of their cellular counterparts, where basically studies have been done in some specific viruses such as those infecting the archaea *Sulfolobus* genus at high temperature and low pH springs (Bautista et al., 2017; Held and Whitaker, 2009; Snyder et al., 2007). The local scale study of Snyder et al., in 2007 included three hot springs in YNP that showed a highly diverse viral community of Sulfolobus islandicus rod-shaped viruses (SIRVs) belonging to family Rudiviridae and Sulfolobus spindle shaped viruses (SSVs) belonging to family Fuselloviridae, that were not limited by dispersion between those springs (Snyder et al., 2007). On an intercontinental scale, four hot springs geographically isolated from North America and Russia were investigated by Held and Whitaker in 2009, where Sulfolobus fuselloviruses (SSVs) showed a biogeographic pattern similar to that of their host (Held and Whitaker, 2009). These authors proposed a plausible model for biogeographic patterns observed in viruses within hot springs, where

virus high migration without their host leads to local adaptations that are stronger than the effects of gen flow between distant springs trough new migrations(Held and Whitaker, 2009).

More recently, a more extended study in Sulfolobus rudiviruses (SIRVs) from 24 acidic hot springs in seven different regions of YNP and five hot springs from Kamchatka (Russia) showed that SIRVs have a biogeographic distribution on a global and scale and also on a local scale within YNP (Bautista et al., 2017).

Although at first glance, there are controversial results between studies, these could be due to methodological differences such as the use of a single gene that might not be enough to resolve the viral spatial structure or the biogeographic patterns associated with these viral communities (Bautista et al., 2017; Held and Whitaker, 2009).

In order to better understand these viral biogeographic patterns and the influence of ecological drivers, such as pH and temperature on their structure, we have analyzed all the viral metagenomes available in public databases to date from different hot springs in Africa and North America, as well as three new viral metagenomes obtained from South America (Chile). We have used the current state of the art database independent approaches, such as *k-mer* frequencies, viral Protein Clusters (vPCs) and viral OTUs (vOTUs), to catalog and compare the space of viral sequences present in all these distant hot springs. These analyses unveiled viral community biogeographic patterns, suggesting that viruses can be passively transported by air, at a local scale and probably on a global scale, to get then locally structured by environmental conditions (pH and temperature) that affect the host community structure.

2. MATERIAL AND METHODS

Patagonia Sampling site.

Porcelana (42º 27’ 29.1’’S-72º 27’ 39.3’’W) and Cahuelmó (42º 15’ 11.8’’S-72º 22’4.4’’W) hot springs are located in Chilean Patagonia. The linear distance between Porcelana and Cahuelmó hot springs is 25.4 km. Porcelana hot springs have a (circum)neutral pH range 7.1-6.8 and temperatures ranging from 60 ºC to 46 ºC, when sampled on December 2014. Cahuelmó hot springs have an alkaline pH range 9.5-9.1 and temperatures ranging from 62 ºC to 38 ºC, when sampled in the same date.

Viral particles enrichment and purification from interstitial fluid from Patagonia microbial mats.

Viral communities from phototrophic microbial mats growing in ponds, where surface water reached 50 or 55 ºC in Porcelana and 51 ºC in Cahuelmó were sampled at noon (12:00 PM) as described in (Guajardo-Leiva et al., 2018). Briefly, five liters of interstitial fluid was squeezed using 150 µm sterilized polyester net SEFAR PET 1000 (Sefar, Heiden, Switzerland), transported in dark blue PET drums and stored at 4 ºC until serial filtration through 0.8 µm pore-size polycarbonate filters (Isopore ATTP, 47 mm diameter, Millipore, Millford, MA, USA) using a Swinex filter holder (Millipore) and 0.22 µm pore-size (Sterivex PES, Millipore). Particles in the 0.22 µm filtrate were concentrated to a final volume of approximately 35 ml using a tangential-flow filtration cartridge (Vivaflow 200, 100 kDa pore size, Vivascience, Lincoln, UK).

Purification of viral particles was done using CsCl density gradient ultracentrifugation as described in (Thurber et al., 2009). Briefly 8 mL of virus enriched interstitial fluid was brought up to 1.12 g mL-1 with CsCl and then loaded onto the heavier CsCl layers of 1.7, 1.5, 1.35 and 1.2 g mL-1 CsCl prepared using SM buffer (NaCl 100mM, MgSO4 8mM and Tris-Cl 50 mM pH 7.5). Centrifuged at 21,755 RPM at 4ºC for 2 h in a swinging bucket rotor Beckman SW40Ti. DNase treatment of the viral CsCl fraction was used to remove the remaining free DNA using 300U of DNAse I for each 1 mL of sample.

DNA extractions and high throughput sequencing of Patagonia viral communities.

Viral DNA was extracted as previously described in (Thurber et al., 2009) by formamide/cetyltrimethylammonium bromide (CTAB) method, followed by phenol/chloroform method. Quality and quantity of the extracted nucleic acids were checked by 1% agarose electrophoresis and fluorometric quantitation using a Qubit (Invitrogen, Waltham, Massachusetts, USA) to be kept at -80 °C until use. Bacterial DNA contamination was checked by 16S rRNA gene PCR amplification using bacterial universal primer (357F CTCCTACGGGAGGCAGCAG and 907R CCGTCAATTCMTTTRAGTTT) as described in (Mackenzie et al., 2013).

Viral DNA samples were sequenced by Illumina Mi-seq technology (Roy J. Carver Biotechnology Center, Illinois, USA). Briefly, shotgun viral DNA libraries were prepared with KAPA Hyper Prep kit (Kapa Biosystems, Wilmington, Massachusetts, USA). Libraries were pooled, quantified by qPCR and

sequenced on one MiSeq flowcell for 251 cycles from each end of the fragments, using a MiSeq 500-cycle sequencing kit version 3 (Illumina, San Diego. California, USA).

For quality filtering, the following filters were applied using Cutadapt, (Martin, 2011) leaving sequences longer than 30 bp (-m 100), with a 3' end trimming for bases with a quality below 30 (-q 30), a hard clipping of the first 9 leftmost bases (-u 9), and finally a perfect match of at least 10 bp (-O 10) against KAPA Hyper Prep for Illumina adaptor. Finally, the removal of sequences representing simple repetitions, that are usually due to sequencing errors, was applied using PRINSEQ (Schmieder and Edwards, 2011) DUST threshold 7 (-lc_method dust, -lc_threshold 7). Details of the number of sequences obtained are shown in Supplementary Table S1.

Public and Patagonia Viral metagenomes assembly and gene prediction.

Public available viral metagenomes from different physicochemical properties (such as pH and temperature), geographic locations (longitude and latitude) and sample source (water, sediment, microbial mat) were found on Virome web-application (Eric Wommack et al., 2012) and downloaded from the imicrobe website (www.imicrobe.us). Accession numbers and links are listed in Supplementary Table S2.

Then the viral metagenomes, both public and also those generated in this study, were assembled using De Bruijn graphs as implemented in the Spades assembler (Bankevich et al., 2012) or MEGAHIT (Li et al., 2016) in metagenomic mode.

A co-assembly strategy was followed for vOTUs related analyses, using quality filtered reads from different samples of the same hot spring as the case of Porcelana, Nymph Lake and Octopus.

An In-silico decontamination of cellular and phiX-174 sequences was performed trough mapping of contigs ≥ 500 pb to bacterial, archeal and phiX-174 sequences contained in RefSeq release 86 using BLASTN (Camacho et al., 2009) (-evalue 0.00001) and query coverage ≥ 5%. Protein-coding gene predictions was made using Prodigal software (Hyatt et al., 2010) options (-p meta -n) only for contigs ≥ 500 pb.

Reference dependent analyses: Taxonomic assignation and abundance quantification.

For taxonomic assignment, predicted proteins from contigs ≥ 500 pb were aligned against the NCBI nr database using DIAMOND (Buchfink et al., 2014) (evalue 0.00001) and parsed using the lowest common ancestor algorithm trough MEGAN 6 (Huson et al., 2016) (LCA score =50).

To quantify the abundance of mapped proteins, reads recruitment from each viral metagenome was performed using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1) and resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Relative abundances were normalized by gene length and library size of each viral metagenome.

Reference independent analyses: MinHash distance, Protein Clusters and Viral vOTUs.

Pairwise mutation distance was calculated for predicted proteins and genes on contigs ≥ 500 pb in all samples using MinHash dimensionality-reduction technique implemented in MASH (Ondov et al., 2016) options (-k 16 and -s 1000000) for genes and (-k 7 and -s 1000000) for proteins.

Predicted proteins ≥ 60 aa in contigs ≥ 500 pb were used to form protein clusters (vPCs) between all viral metagenomes as described on (Brum et al., 2015; Yooseph et al., 2007). Proteins were initially mapped to existing vPCs from hot spring ecosystem type on IMG/VR database, by using CD-HIT-2d (Li and Godzik, 2006) at 60% of identity and 80% of coverage. Then, the remaining, unmapped proteins were self clustered by CD-HIT (Li and Godzik, 2006), using the same options as above. After clustering step all vPCs are quantified in each viral metagenome using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1) and resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Relative abundances were normalized by gene length of each cluster and library size of each viral metagenome as described in (Yooseph et al., 2007; Brum et al., 2015).

vPCs were functionally annotated using Pfam database through hmmscan options (-cut_ga) implemented on HMMER3 (Eddy, 2009) using representative sequences of each vPCs.

vPCs normalized abundance are used to calculate alpha functional diversity by Shannon H index and evenness by Pielou´s index. Also functional beta diversity, Bray-Curtis distance was calculated using the Vegan package (Oksanen et al 2019).

Viral vOTUs are constructed in basis to contigs ≥ 5 kb from all samples clustered with nucmer algorithm implemented in MUMmer3 (Kurtz et al., 2004) ≥95% identity and ≥ 80%coverage, as in (Duhaime et al., 2017; Paez-espino et al., 2019; Roux et al., 2017), to generate a pool of non-redundant population contigs.

Each vOTU was quantified in each viral metagenome using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1) and resulting SAM file was parsed by BBmap pileup script (Bushnell B. - sourceforge.net/projects/bbmap/). Only if ≥ 75% of the vOTU sequence length is covered, the vOTU was considered as present in the sample. Relative abundances of viral vOTUs were normalized by Trimmed Mean of M-values (TMM) algorithm, implemented in EdgeR package (Robinson et al., 2009) and vOTU sequence length as in (Roux et al., 2017). Normalized abundances are used to calculate species alpha and beta diversity by Shannon H index, evenness by Pielou´s index and Bray-Curtis distance using the Vegan package (Oksanen et al., 2019).

Statistical Analysis

Relationship between hot springs properties (pH and temperature) and the changes in alpha-diversity (Shannon index) and evenness (Pielou's), was tested using linear models (linear and quadratic regressions) in R.

Permutational multivariate analysis of variance (PerMANOVA, "adonis2" function in vegan R package) (Oksanen et al., 2019) was used to test the main effect of hot spring properties (pH, temperature, sample source and geographic location) listed on Table 1, over a Bray-Curtis distance matrix of vPCs and viral vOTUs with 9999 random permutations.

Cluster analysis was performed on Bray-Curtis and MASH distance matrices of proteins, genes, vPCs and viral vOTUs by using an unweighted pair group mean (UPGMA) algorithm implemented in the 'hclust' function of vegan R package (Oksanen et al., 2019). Bray-Curtis distance matrices were further analyzed by Principal component analysis (PCoA) plot, using ampvis2 R package (Albertsen et al., 2015). To detect possible associations between viral community structure and hot springs physicochemical properties (supplementary variables), the vectors of significant hot springs physicochemical factors ($P < 0.05$) and properties were fitted onto PCoA ordination space using the 'envfit' function of the vegan R package (Oksanen et al., 2019) with 999 random permutations.

Host assignment through CRISPR spacers.

Assemblies of Patagonia cellular metagenomes (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Guajardo-Leiva et al., 2018), were taxonomically grouped into Metagenome Assembled Genomes (MAGs), using the Expectation-Maximization (EM) algorithm implemented in MaxBin 2.0 (Wu et al., 2016). In order to assess the completeness and contamination of each MAGs, CheckM analyses (Parks et al., 2015) were performed. Finally, the closest genome of each bin was searched using the Tetra Correlation Search (TCS) analysis implemented in Jspecies tool (Richter and Rossello-Mora, 2009) with selection criteria of Z score greater than 0.999 and ANI ≥ 95%(Konstantinidis et al., 2017).
CRISPR loci were identified in MAGs from Patagonia cellular metagenomes using CRISPRFinder tool (Grissa et al., 2007). Spacers from CRISPR containing contigs were mapped to vOTUs set using Bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1).

Monopartite network analysis of viral communities.

Protein monopartire networks implemented in vContact (Bolduc et al., 2017) were constructed using vOTUs set (considered here as viral genomes) from all samples. Briefly, predicted proteins from vOTUs were compared by BLASTP (Camacho et al., 2009) in an all-versus-all pairwise comparison (-evalue 0.00001, bitscore 50). Protein clusters were subsequently identified using the Markov clustering algorithm (MCL) based on BLASTP e-values with an inflation value of 2, building protein cluster profiles for each genome and generating a similarity network.
Sequences from hot springs present in IMG/VR database (Paez-Espino et al., 2017) with information about putative hosts (144 contigs) were used as references in the network construction.
For network visualization we used an edge-weighted spring embedded model implemented in Cytoscape 3.6.1 (Shannon et al., 2003), which places the genomes sharing more proteins closer to each other conforming viral modules. Viral modules were then organized according to their predicted host from our previous analysis, BLASTN against viral RefSeq release 91 (-evalue 0.0000000001) and query coverage ≥ 51% or reference information according to IMG/VR database.

3. RESULTS

Hot springs used in this study can be divided into four main categories according to their pH and temperature (Table 1). First, we found a group of acidic (pH 2.0-3.6) and high-temperature hot springs (80-86 °C) from YNP such as Nymph Lake (NL) and Crater Hills Geyser (CHAS). A second, group was formed by circumneutral pH (7.3-8.1) and high-temperature hot springs (74-93 °C) also from YNP Low Geyser Basin (LGBs) such as Octopus and Bear paw. The third group was composed of circumneutral to alkaline pH (5.7-9.4) and mild-temperature hot springs (50-60 °C) from South America (Patagonia, Chile) such as Porcelana and Cahuelmó, and from Brandvlei in South Africa. Finally, a fourth category was formed by the only acidic (pH 2) and mild temperature (52 °C) hot spring from California as it is Boiling Springs Lake.

Viral metagenomes assemblies.

The sequencing deep of public available viral metagenomes (Supplementary Table S1), varies enormously depending on the sequencing technology used (8.4 to 870 $^{10^3}$ sequences), however, this is partially compensated by the greater length of the sequences obtained in those with less depth. Illumina viral metagenomes from Patagonia allowed us to obtain a high number of sequences (Supplementary Table S1), becoming the viral metagenomes of hot springs with the highest number of bases available to date (2.3 to 4.3 $^{10^6}$ reads).

Assembly of the different viral metagenomes yield and uneven number of contigs ≥500 pb, ranging from 1758 to 16894 contigs (Supplementary Table S1) using 33-100% of the reads. The number of predicted proteins (Supplementary Table S1) showed a high dispersion (2815 to 40845) and depend directly on the contig numbers in each sample.

Database dependent analyses.

Predicted proteins on contigs ≥500pb were used for a reference based analyses using NCBI nr database. Proteins that map to NCBI nr database, represent 4 to 87% of total reads in each sample (Supplementary Table S1) evidencing that for most of the samples, predicted protein are unknown and absent in current databases. Classification of predicted proteins at the Domain level (Figure 1A)

showed that 33 to 98% of the mapped sequences were classified as proteins of cellular origin. Patagonia (Cahuelmó and Porcelana), South Africa (Brandvlei) and YNP Octopus G7162 sample had most of the cellular sequences (91 to 62%) classified as bacteria. On the other hand, cellular sequences from most North America samples such as Bear paw, Crater Hills Geyser (CHAS), Nymph Lake (NL10, NL17 and NL18), Octopus and Boiling Spring Lake (BSL), were mostly classified as archaea (11 to 71%). Proteins classified as eukaryotic origin were only abundant (40% and 19%) in CHAS and NL18 samples from YNP, respectively.

Proteins classified as viral origin (Figure 1A) shown a high dispersion in their relative abundances, representing 6 to 64% of the total mapped proteins. Generally, most of the viral proteins in samples CHAS, NL10, NL17, NL18 and BSL were classified as archaeal viruses belonging to families Rudiviridae (28 to 40%), Lipothrixviridae (1 to 13%), Bicaudaviridae (1 to 11 %), well known to infect Sulfolobales, together with some unclassified archeal dsDNA viruses (23 to 44%). Meanwhile samples from Bear paw and Octopus, showed also that most of their viral proteins were classified as archaeal viruses but this time associated to the Hypherthermophilic Archaeal Virus 1 (HAV1) (70 to 85%), and to the Globuloviridae family (10 to 20%) that infect Thermoproteales order members.

On the other hand, viral sequences from Patagonia and South Africa samples were classified within known viral families inside the Caudovirales order that mainly infect bacteria. From those, Podoviridae family was the most represented (19 to 77%) together with unclassified bacterial viruses that were particularly represented (72%) in the Brandvlei sample.

There are some exceptions to these general observations, such as Octopus G7162 sample that has a great contribution of unclassified environmental viruses (mostly bacterial viruses) and a higher ratio of Globuloviridae:HAV1 than Bear paw and Octopus samples. Also, BSL has a high abundance of environmental viruses sequences and the highest number of ssDNA viral sequences. Finally, CHAS has the higher abundance of Siphoviridae sequences which are known to infect both, bacteria and archaea.

Reference-free genetic distance estimation of hot springs viral communities.

Gene and protein sequences (161391) predicted on contigs longer than 500 pb were used for genetic distance estimation, using MinHash dimensionality-reduction technique, implemented in MASH. UPGMA clustering analysis based on MASH genetic distance, of *k-mer* frequencies of gene and

protein sequences from hot springs viral metagenomes showed three distinct clusters of samples, found at both the nucleotide or aminoacid level (Figure 2).

The first cluster was formed by NL (NL10, NL17, NL18) and CHAS samples from YNP with acidic pH ranges between 2 to 3.6 and temperatures between 80 and 86 ºC, hereafter referred as YNP Acidic cluster. The second cluster was formed by samples from the LGB (Bear paw, Octopus, and Octopus G7162) also from YNP, with circumneutral pH ranges between 7.34 to 8.14 and temperatures between 74 and 93 ºC, hereafter referred as YNP Neutral cluster. Finally, the last cluster was composed by samples from Patagonia hot springs (P50, P55 and CA), with pH ranges between 6.67 to 9.37 and temperatures between 50 and 55 ºC, hereafter referred as Patagonia Phototrophic cluster. Samples from BSL and Brandvlei were part of YNP Acidic cluster and Patagonia Phototrophic cluster respectively, but in both cases were the most distant samples on the group.

Diversity comparison of hot springs viral Protein Cluster (vPCs).

Proteins ≥ 60 aa (133239 proteins) were used to generate viral protein clusters (vPCs) using CD-HIT. A total of 22667 vPCs with two or more proteins were obtained, representing 4 to 92% of the total reads (Supplementary Table S1) and in most of the samples represented an improvement compared with reference based analyses. From all vPCs, 54 clusters contained from 32 to 10 proteins, 6243 clusters contained nine to two proteins and 15623 clusters have only two proteins. Most of the vPCs were new (93%), with only 1477 vPCs previously reported as part of a hot spring ecosystem type on IMG/VR database.

Viral protein clusters (vPCs) provide a metric of viral communities diversity, through quantification of protein families diversity per sample (Supplementary Figure S1). The calculated alpha diversity (Shannon's index) of vPCs presented an average of 5.47 ± 1.46, where three samples (NL10, NL17, and NL18) had a diversity above this average, and one sample (Cahuelmó) below this average. Total vPCs diversity for the pool of normalized vPCs obtained from all samples was 7.6, which is above the average, but closer to the most diverse sample NL10 (7.55). Viral PCs calculated evenness (Pielou's) represented an average of 0.74 ± 0.17, with only two samples (NL10 and NL17) above the average, and one sample (Cahuelmó) below the average. Total vPCs evenness from all samples was 0.76, which is within the average of each individual hot spring sample.

Linear models were used to test the relationship of environmental parameters such as pH and temperature with vPCs diversity and evenness. The analyses shows a moderate correlation (Supplementary Figure S2) between diversity and evenness with pH (r=0.55 and r = 0.67 respectively) and temperature (r = 0.53 and r = 0.6 respectively), where pH was considered a better predictor of diversity and evenness for hot springs viral proteins.

Differences in vPCs between samples allowed for determination of the viral compositional dissimilarities degree and its relation to environmental conditions, through beta-diversity estimation (Bray-Curtis dissimilarity). Hierarchical clustering of hot springs vPCs based on Bray-Curtis distance (Figure 3A) showed three clusters of samples, which were correlated to the same clusters observed in the previous analyzes of genetic distance using MASH (Figure 2). The first group was represented by YNP Acidic cluster which was conformed by NL (NL10, NL17, NL18) and CHAS samples, the second group corresponding to YNP Neutral cluster formed by LGB samples (Bear paw, Octopus and Octopus G7162), and finally the third group was represented by Patagonia Phototrophic cluster that was composed by Patagonia hot spring samples (P50, P55 and CA). Again, samples BSL and Brandvlei were the most dissimilar samples.

Analyses of vPCs distribution between clusters of samples (Supplementary Figure S3) corroborate the correct assignation of the group of samples found here and also on the previous MASH analyses, showing a minimal number of shared PCs between the three different groups of samples.

Principal Coordinates Analysis (PCoA) based on hot springs vPCs Bray-Curtis distances (Figure 3B), also corroborate the separation of viral communities in three different clusters. The first two axis of the PCoA explained about 36% of the total variance, with axis 1 and 2 explaining 20.6% and 15.3% of the total variance, respectively. Vectors of environmental variables were fitted into the ordination space, to assess the effect of pH and temperature physicochemical properties, geographic location (longitude and latitude) and sample source (water, sedimet, microbial mat). The environmental fitting analysis indicated that all these hot spring properties except for the longitude were strongly correlated with viral communities ordination ($P < 0.05$, 9999 permutations), where the length of the arrow vector directly correlate with the fitness of the predictor. Additionally these same factors (pH, temperature, latitude and sample source) were used in a PerMANOVA analyses (Table 2) under a reduced model (Type III, marginal test) indicating that pH is the main explanatory factor (explaining ~12%) forthe total variance of viral community beta diversity ($P < 0.0735$, 9999 permutations).

Functional annotation of vPCs using Pfam database, allows to identify 4826 vPCs that codified 1364 different protein families (PFs). From these 1364 PFs, 1361 were grouped into the three main hot springs cluster (YNP Acidic, YNP Neutral and Patagonia Phototropic). Venn diagram (Figure 4) showed that 42 PFs represented core proteins that are shared for all hot springs clusters where viral communities from Patagonia hot springs mats concentrate the largest number (1081) of exclusive PFs. Hot springs core PFs are not restricted to typical viral hallmark genes and includes proteins of unknown functions (DUF) and cellular proteins related to DNA and protein metabolism (Supplementary Table S3).

Abundance analyses of the 1364 PFs across all samples, showed that only 27 of these PFs were ≥1% of abundance (Supplementary Table S4), but all together represented 45% of the total PFs abundance.

Diversity comparison of viral Operational Taxonomic Units (vOTUs).

Viral OTUs are defined as population consensus genomes or quasi-genomes, constructed by non-redundant viral contigs ≥ 5000 pb. Here using 1162 contigs a set of 948 vOTUs were obtained. Most of the contigs (814) were unique for each sample and 348 contigs formed 134 groups within an ≥95% identity and ≥80%coverage thresholds. Detection of vOTUs was possible in 9 of 12 samples (except for Bear paw, Octopus and BSL), considering a 75% threshold on the contig length coverage.

Patagonia Phototrophic cluster concentrated the major number of exclusive vOTUs (857), followed by Brandvei (55), YNP Acidic cluster (28) and Octopus G7162 (8). Most of the vOTUs were shared only inside each cluster (Patagonia Phototrophic and YNP Acidic). Inside the YNP Acidic cluster, 21 vOTUs were shared by all NL samples (NL10, NL17 and NL18) and only 7 vOTUs by all samples. On the other hand, Patagonia Phototrophic cluster showed 349 shared vOTUs, but also many exclusive vOTUs were found for each sample (68, 352 and 89 for CA, P50 and P55 respectively) (Supplementary Figure S4). Additionally, Brandvlei shared a unique vOTU with P50.

Alpha diversity (Shannon's index) and evenness (Pielou's) were calculated through the quantification of the relative abundance of each vOTU in each sample (Supplementary Figure S5). Shannon's index had an average of 3.22 ± 1.26, finding three samples (CA, P50 and P55) with a diversity above the average, and two samples from YNP (CHAS and Octopus G7162) were below the average. Total

diversity for hot spring viral communities (Shannon index = 4.92) was obtained from the pool of normalized vOTUs of all samples, and estimated to be above the average, but closer to the most diverse sample P55 (4.76). Pielou's evenness had an average of 0.81 ± 0.17, finding one sample (CHAS) above and one sample (Octopus G7162) below the average. Total evenness from all samples was 0.72 which is inside the range of evenness for individual samples. When linear models were used to test the correlation of vOTUs diversity and evenness with temperature, a strong to moderate correlation was found (Supplementary Figure S6) (r = 0.85 and r = 0.69 respectively). For pH a weak to moderate correlation was found (r=0.38 and r = 0.58 respectively). Then the temperature was pointed to be the best predictor of diversity and evenness for hot springs viral communities.

Beta diversity estimation of hot springs viral communities based on vOTUs information, showed similar results than vPCs analysis. Hierarchical clustering of vOTUs based on Bray-Curtis distance (Figure 5A) separate the same two clusters of samples observed in the previous analyzes using MASH (Figure 2) and vPCs (Figure 3). The first cluster was conformed by NL (NL10, NL17, NL18) and Crater Hills Geyser (CHAS) samples, while the second cluster was formed by viruses from Patagonia hot springs (P50, P55 and CA). On the other hand, viral communities from Octopus G7162 and Brandvlei were highly dissimilar and they did not form part of any cluster.

Principal Coordinates Analysis (PCoA) based hot springs vOTUs Bray-Curtis distance (Figure 3B) corroborate the organization of viral communities in two main clusters. The first two axis of the PCoA explained about 48% of the total variance, with axis 1 and 2 explaining 31.1% and 16.9% of the total variance, respectively. Vectors of environmental variables were fitted into the ordination space, to asses the effect of physicochemical properties (pH and Temperature), geographic location (longitude and latitude) and sample source (water and microbial mat). The environmental fitting analysis indicated that all properties except for longitude were strongly correlated with viral communities ordination ($P < 0.05$, 9999 permutations) where the length of the arrow vector directly correlate with the strong of the predictor.

PerMANOVA analyses under a reduced model (Type III, marginal test) using the correlated factors obtained in the PCoA, indicated that none of the factors analyzed have a main effect over the total variance of the viral community beta diversity (Table 3).

Protein-sharing monopartite networks of hot springs viral communities.

In order to generate a monopartite network representation of the relationship between hot springs viral genomes we used the 948 vOTUs set and 144 contigs from the IMG/VR database that contained host information, resulting in a total of 1092 viral genomes analyzed.

The resulting network (Supplementary Figure S7), that represents statistically significant relationship between the protein profiles of the viral genomes, consisted in 784 genomes (nodes) and 1896 relationships (edges). The network, showed 233 connected components (viral clusters or modules), where 67% of the modules (157) were formed by only two genomes. Other simple parameters of the network such as clustering coefficient, network diameter, radius, density and heterogeneity are showed in Supplementary Table S5.

A subset of the network (that contains 105 new viral genomes or quasi-genomes and 121 genomes from IMG/VR database) was selected to explore the taxonomy and host prediction of the hot springs viral genomes (Figure 6). The 226 total nodes meet at least one of the following three criteria. First, to have a match to predicted host CRISPR spacer(s). Second, to have BLASTN best hit under strict parameters (e.g. e-value 10-10). Third, to be part of a module that contains at least one node that meets either the first or second criteria. This allowed us to identify 39 Viral Clusters (VCs) with their respective host prediction and/or taxonomy. These 39 VCs could infect 24 different taxa of bacteria and archaea classified at species (19), order (3) or phyla (2) levels. Most of the VCs are made up from viral genomes with a single host genus, a unique exception is the case of two VCs that would infect the bacterial genera Roseiflexus and Meiothermus. Groups with broader hosts (Phyla and Order) are formed mostly by IMG/VR genomes, where host annotations were broader.

Taxonomic prediction was possible for eigth of the VCs, two groups were assigned to Podoviridae family, infecting the bacterial genera Fischerella and Rhodoferax, one group was assigned to Siphoviridae family infecting Termoanaerobacteriales members, one group was assigned to Sphaerolipoviridae family infecting Thermus and three groups were assigned to Lipothrixviridae, Rudiviridae and Turriviridae families infecting Sulfolobales (Sulfolobus and Metallosphaera).

Two-thirds of the VCs were composed only by genomes recovered from the same hot spring, with some exceptions of modules with genomes from Porcelana-Cahuelmó, Brandvlei-Porcelana-Cahuelmo and IMG/VR with different hot springs.

4. DISCUSSION

Hot springs are discontinuous environments that can be considered as "hot islands" surrounded by a "cold ocean" providing a unique study model in which to evaluate the relevance of environmental factors and dispersal limitation in the establishment and development of viral communities. Analyses of viral communities by *k-mer* frequencies, protein clusters (vPCs) and populations (vOTUs) revealed biogeographic patterns, suggesting that viruses can be passively transported by air, on local scale but probably also on a global scale and then locally structured by environmental conditions (pH and temperature) that affect the host community structure. This is emphasized by the analysis of monopartite networks, which shows that the presence of specific viral genus or families or both, directly depends on the existence of related hosts. The last, would disrupt the gene flow of viral populations according to the environmental conditions that limit the colonization and growing of microbial hosts.

Insights into viral communities structure in hot springs.

It is well known, that pH and temperature allows for an important distinction between microbial communities inhabiting hot springs (Inskeep et al., 2013; Power et al., 2018). Colonization by microbial phototrophic mats not only has temperature limits but also are dependent on pH. The upper temperature limit estimated for oxygenic phototrophs is near to 74˚C (Brock, 1973), but Cyanobacteria (the main oxygenic phototrophic phyla in many hot springs mats) generally do not live in habitats below pH 4.5-5. Consistently, there is also a phototrophic limit at pH values < 5, and temperature near to 56˚C that correspond to the upper temperature limit for other microbes such as Cyanidales, diatoms, or photosynthetic proteobacteria (Inskeep et al., 2013).

Morover, pH also produces dichotomous divisions in hyperthermophilic (72-98 °C) environments. Sites of pH values >4, are dominated by bacteria from Aquificales and Thermotogales, with contribution of archaea from Desulfurococcales and Thermoproteales. However, at pH values <4, Sulfolobales archaea

553 dominate (Inskeep et al., 2013; Menzel et al., 2015). According to the above, it is expected that viral
554 communities at different pH and temperature, have a structure concordant with the hosts that inhabit in
555 each particular hot springs (Gudbergsdóttir et al., 2016).
556
557 Our database dependent analyses (Figure 1A), exactly reflects the proposed scenario where dominance
558 of bacterial or archaeal viruses occur according to the structure of their host community. Even when at
559 a first look 44 to 94% of the viral proteins obtained from the 12 metagenomes here investigated were
560 classified as cellular origin, which is not uncommon in viral metagenomics studies (Edwards and
561 Rohwer, 2005; Roux et al., 2012), those proteins were assigned to the most represented domain
562 (Bacteria or Archaea) in each sample. The high number of viral proteins assigned to the cellular taxa
563 can be partially explained by the lack of viral gene annotations in databases, but also by the potential
564 HGT between viral and host genomes (Edwards and Rohwer, 2005; Roux et al., 2012). The first
565 explanation is especially evident in samples rich in bacterial viruses, since only 17 representatives of
566 thermophilic viral sequences are present in the databases. It is not the same case for samples rich in
567 archaeal viruses, since most of the hot springs related viral sequences that exist in databases correspond
568 to those that infect archaeal hosts (Zablocki et al., 2018).
569
570 Viral proteins assigned to viruses also separated hot springs samples dominated by archaeal viruses
571 from those of bacterial viruses (Figure 1B). As in previous studies, in Nymph Lake and also Crater Hill
572 Geyser (Bolduc et al., 2012, 2015) low pH-high temperature hot springs, there was an enrichment of
573 archaea infecting viruses sequences from Rudiviridae, Lipothrixviridae and Bicaudaviridae families,
574 which infect members of the Sulfolobales (Bolduc et al., 2012, 2015; Menzel et al., 2015).
575 Meanwhile, circumneutral pH-high temperature sites, were dominated by both, unclassified
576 environmental bacterial viruses and archaeal viruses related to Globuloviridae family and the
577 Hypherthermophilic Archaeal Virus 1 (HAV1) which infect Thermoproteales order, similar to findings
578 from previous studies (Pride and Schoenfeld, 2008; Schoenfeld et al., 2008). The last group of hot
579 springs with circumneutral pH but low temperatures, were rich in bacterial viruses from Podoviridae
580 family and unclassified bacterial viruses. The high abundance of Podoviridae sequences was especially
581 interesting since there is a few studies that have reported this family in geothermal environments
582 (Guajardo-Leiva et al., 2018; Zablocki et al., 2017, 2018).

In this study, BSL was the only representative of low pH-low temperature hot springs, consequently it was not surprisingly that it has a unique combination of viral populations (Figure 1B). This sample was rich in archaeal viruses from Rudiviridae, Bicaudairidae and Turriviridae families, but also in unclassified ssDNA viruses, environmental viruses such as BSL-RDHV and representatives of Circoviridae family. The Circoviridae related sequences were the only group of viruses previously reported in BSL (Diemer and Stedman, 2012).

Database-independent clustering of hot springs viral communities

Quantitative measurements of variation in community structure among samples over spatial, temporal or environmental gradients have been an endless task in microbial ecology. Usually, this work has been done based on universal molecular markers to calculate ecological distances using taxonomic or phylogenetic information (Jiang et al., 2012; Maltez Thomas et al., 2018). However, in viral communities, the absence of universal molecular markers, as well as a diffuse taxonomy and phylogeny, have made it difficult to quantify the variation in the structure of these communities.

Quantification of *k-mer* frequencies has been used widely for the comparison analysis of large DNA sequences such as chromosomes, whole genomes, and metagenomes (Hildenbrand et al., 2017; Jiang et al., 2012; Ondov et al., 2016). Lately, *k-mer* analyses of dsDNA microbial viruses using MASH (Ondov et al., 2016) have shown that *k-mer* comparison analyses can yield similar results than alignment-based ANI comparisons (Mavrich and Hatfull, 2017). Then, we chose the use of database independent methods such as the quantification of *k-mer* frequencies in viral proteins and genes to compare the structure of viral communities from hot springs with different physicochemical conditions and geographic locations (Figure 2). This, under the rationale that there exist a high structure conservation (within a broad phylogenetic spectrum), within the coding DNA sequences (CDS) of all known life forms (including viruses), due to the information storage capacity of the chemical structure of proteins as amino acid codons (Sievers et al., 2018).

As mentioned before, environmental conditions (pH and temperature gradients) defined and separate the different types of hot springs. This division is coincident with the cluster separation (YNP Acidic, YNP Neutral and Patagonia Phototrophic) of the viral populations found by MASH genetic distance analyses. Therefore, the environmental conditions at sites seem to be relevant for the establishment and development of the potential hosts, harboring the viral communities of these three clusters and

consequently determining the final structure of the viral community, which could potentially explain the genetic discontinuity that has been observed in hot spring viruses (Duffy et al., 2007).

It is important to note, that even when amino acids evolve slowly and are characters subject to convergence, they provide better resolution in analyses of viral populations by *k-mer* frequencies (see scale of Figures 2A and 2B) than nucleotide sequences, probably due to the adverse effects of the base composition and the codon preference biases that affect the genes (Jun et al., 2009).

Role of physicochemical factors in the universe of hot springs viral proteins.

The exploration of the protein universe begins by grouping proteins into clusters or families of evolutionary related sequences. The methodology used for this task typically relies on sequence identity to be grouped into families based on conserved structural units (domains) or according to the full-length sequences (Yooseph et al., 2007; Zaslavsky et al., 2016). Therefore, we can define protein clusters as groups of homologous proteins that probably share the same or similar function (Yooseph et al., 2007). We used this methodology as an approach to calculate the diversity (Shannon Index) and evenness (Pielou's) of viral communities (Suplementary Figure S1). Diversity (5.37 ± 1.46) and evenness (0.74 ± 0.17) calculations using vPCs resulted similar to previous estimations of diversity on YNP viral metagenomes using PHACCS (Angly et al., 2005) predictions (Bolduc et al., 2015; Schoenfeld et al., 2008) such as Nymph Lake (Shannon: 5.37 ± 0.72 and Pielou's: 0.91 ± 0.024), Bear Paw and Octopus (Shannon: 6.23 ± 0.67 and Pielou's: 0.94 ± 0.008). Although both methodologies estimate diversity differently, the results were concordant, indicating that these estimates probably approximate to the true viral diversity of these systems.

Furthermore, the moderate correlation found for the alpha diversity and evenness with pH ($r=0.55$ and $r = 0.67$ respectively) and temperature ($r = 0.53$ and $r = 0.6$ respectively) show that both physicochemical factors have a predictive effect on diversity and evenness of viral communities in these hot springs (Supplementary Figure S2). These findings are in agreement with the cellular counterpart diversity reports from hot springs around the world, where a moderate correlation with pH (0.4-0.44) (Power et al., 2018; Sharp et al., 2014) and a stronger correlation with temperature (0.79) (Sharp et al., 2014), was found.

643 Therefore, and even with the obvious limitations of the reduced number of hot spring viral
644 metagenomes available for analyses, our results strongly suggest that viruses have diversity and
645 evenness patterns that are affected by pH and temperature as well as their host counterpart.
646
647 Clustering of viral populations through vPCs, deepen in the notion of restrictions to gene flow between
648 hot springs, even in those sites that have similar physicochemical properties and closer geographic
649 locations. The latter becomes more evident when observing the dissimilarities between samples from
650 the three different clusters (YNP Acidic, YNP Neutral and Patagonia Phototrophic) and also within
651 each cluster.
652
653 Ordination of vPCs (Figure 3B) confirms the results of the hierarchical clustering of viral communities
654 (Figure 3A) even when only ~ 36% of the variance was explained by PCo1 and PCo2. Stronger
655 predictors ($p<0.05$) of the viral structure were the environmental variables pH and latitude, whereas
656 temperature can be considered weaker (because of arrows lengths). Categorical variable, sample
657 source, also have a strong effect on the ordination, particularly on phototrophic microbial mats from
658 Patagonia (P50, P55 and CA) and sediment sample from BSL in California.
659 Permanova analyses (Table 2) corroborate the explanatory role of pH (12% of the total variance) in the
660 dissimilarity of viral communities structure, even when it is not strictly statistically supported (> 90%
661 confidence interval).
662 Together these results showed an organization of the viral communities in the physicochemical
663 gradients across the samples and a different viral protein repertory at each hot spring, suggesting that
664 geographic location (latitude) and pH are the factors that more influenced the viral community
665 structure. In that sense, latitudinal biogeographic patterns have been extensively studied in
666 macroorganism (Hillebrand, 2004) and also been described in marine (Fuhrman et al., 2008) and
667 terrestrial microorganisms (Andam et al., 2016; Sharp et al., 2014). Viral counterpart has also shown
668 latitudinal influence on community structure in marine datasets from the Pacific Ocean Virome (POV)
669 (Hurwitz et al., 2014) and Tara Ocean Virome (TOV) (Brum et al., 2015). However the mechanism
670 underlying these latitudinal patterns are still under debate for viral communities. Here, and due to the
671 nature of vPCs that can measure deep evolutionary events, these patterns are probably the product of
672 historical events (geological, ecological or demographic), such as glaciations, marine currents, and
673 atmospheric circulation that influenced their dispersion and diversification(Andam et al., 2016).

Regarding pH, has been described as a significant factor that influence the microbial community composition in hot springs(Inskeep et al., 2013; Menzel et al., 2015). It has even been proposed as the main driver of these communities to date (Power et al., 2018). It is also known that pH generates specific adaptations in microorganisms to deal with altered nutrient availability, metal solubility and organic carbon characteristics that result in a reduced number of taxa that can physiologically tolerate these conditions (Power et al., 2018). In this way, pH will affect the composition of the host and consequently the viral communities that prey and reproduce in those environments (niche specialization).

Lastly, when the hot spring vPCs functionally profile was examined through Pfam annotation it showed that only 21% of the viral protein universe has a know function associated with a protein family. This unveiled the enormous functional potential of viruses hidden in these extreme ecosystems.

When unique functions were addressed, only 1/3 of the annotations corresponded to non redundant functions which implies that biological functions of vPCs are above the identity and coverage thresholds widely used (Brum et al., 2015; Yooseph et al., 2007). The latter, is evident by the minimum number of shared vPCs between clusters of samples found here (Supplementary Figure S3), which differs significantly with what was found when shared functions were compared.

Although a small group of functions were shared among all viral communities clusters (Figure 4) some of them represented viral hallmark genes (Suplementary Table S3) such as the DNA polymerase B, the phage integrase, the phage terminase large subunit and the caudovirus prohead serine protease. These hallmark genes are shared by different groups of viruses and do not have close homologs in cellular organisms (Koonin et al., 2006). It is also known that hallmark genes have a fundamental role in maintaining the integrity of the virosphere because they connect different families and orders of viruses (Iranzo et al., 2016). Therefore it is not surprising to find them commonly as part of the hot springs core vPCs. In addition, a large number of functions that have homologous sequences in cellular organisms are also widely distributed among viruses such as proteins related to DNA synthesis and replication which for instance have been found previously in YNP viral metagenomes (Davison et al., 2016). The importance of these other common genes is that they arise as potential new viral markers for the study of hot springs viral communities (Davison et al., 2016). Moreover, specific functions shared only inside each viral community clusters are also of interest here, because they can be

considered as signature genes, where their presence-absence pattern is a diagnosis of specific viral populations occurrence (Iranzo et al., 2016).

An example of particular interest, was the finding of three different PF with glycosyl transferase function among the most abundant annotated vPCs and present in all viral community clusters (Supplementary Tables S3 and S4). Bacterial viruses are known to encode two types of glycosyltransferases. Some lytic viruses glucosylate their DNA to protect it from host restriction systems, while some lysogenic viruses express glycosyl transferases during lysogenic conversion affecting the glycome of the host and in particular the host cell serotype (Markine-Goriaynoff et al., 2004). Together, this could mean that for lytic viruses this evasion mechanism of the bacterial immune system is widely distributed in hot springs.

Endemism of hot springs vOTUs is dependent of geographic location and temperature.

Viral OTUs captures evolutionary and ecologically cohesive populations of closely related viral genomes (genotypes), which have no fitness differences in the same niche space (host) (Duhaime et al., 2017). Therefore, vOTUs provide a metric unit to analyze viral communities at genome and population levels.

The lower values of (Shannon Index) diversity (3.22 ± 1.26) but higher (Pielou's) evenness (0.81 ± 0.17) obtained from hot spring viral communities vOTUs data (Suplementary Figure S5) compared with those obtained from vPC approach, is a trend also showed for global ocean viral communities (Brum et al., 2015). It reflects a known issue of vOTU approximation, where absolute diversity is under estimated, because metagenomic assemblies only yield large contigs for the most abundant viral genotypes (Roux et al., 2017) which in turn depends on sequencing depth and community complexity. Then, diversity should be interpreted with care, even though once normalized, sample metrics comparisons are generally robust to differences in community complexity and sequencing depth (Roux et al., 2017).

On the other hand, the strong correlation (Supplementary Figure S6) found for alpha diversity with temperature ($r = 0.85$), and the weak correlation with pH ($r=0.38$) showed that only the first factor have a predictive effect on viral diversity. However, the moderate correlation of temperature and pH (0.69 and 0.58 respectively) with evenness also indicate that temperature might have a better predictive effect

over evenness. As discussed before, a strong correlation of diversity with temperature (0.79) have been found for hot springs microbial communities (Sharp et al., 2014). The latter supports our funding, despite the small number of the existing samples analyzed, which strongly suggest that viral diversity and evenness are affected mainly by temperature. It is also important to note that, when we removed low diversity samples (CHAS and Octopus G7162) for the analysis to test diversity correlation with temperature and pH (data not show), we obtained a strong correlation with both parameters (0.96 and 0.92 respectively).

Viral genomes and populations hierarchical clustering (Figure 5A) analyses, congruently separated viral communities in two (YNP Acidic and Patagonia Phototrophic) of the three clusters defined on our previous analyses based on *k-mer* frequencies and vPCs. The absence of YNP Neutral cluster was related to the impossibility to detect any vOTU in Bear paw and Octopus samples mostly due to the lack of coverage (75 % coverage treshold) even when vOTUs assembled from Octopus (Octopus and Octopus G7162 co-assembly) were available.

Distance between viral communities inside each cluster were higher than in the vPCs analyses, because of the strictness of vOTUs detection thresholds and the fast evolving nature of DNA with respect to proteins. However, the more permissive nature of vPCs based analyzes is necessary for comparative viral genomics at deeper evolutionary times and therefore not relevant to populations and speciation process.

In viral populations, speciation is proposed to follow a parapatric model, which implies the existence of incomplete genetic barriers and substantial opportunities to gene flow (Duffy et al., 2007). Notwithstanding, because of viruses high mutation rate or reintroduction of alleles that induce slow adaptation of emerging new viruses lineage in novel hosts, ranges of the closely related ancestral and evolved viruses stop to overlap (Duffy et al., 2007). The last scenario could lead to the genetic discontinuity between nearby hot springs and even inside the same hot springs over the thermal gradients. This feature was observed in our analyses, which implies that even when gene flow is not totally interrupted, a local adaptation of host and therefore their viruses could produce the actual structure observed.

Ordination of vOTUs (Figure 5B) confirms the results of the hierarchical clustering (Figure 5A) showing the same groups of viral communities, even when just half of the variance was explained by

PCo1 and PCo2. Environmental variables, stronger predictors ($p<0.05$) of the viral structure, were latitude and temperature, where conversely pH was considered weaker (because of arrows lengths). Categorical variable, sample source, also have a predictive effect on the ordination but was diffuse between samples. However, Permanova analyses (Table 3) produced no significant results for a main effect of the four predictor factors that we found in our exploratory PCoA analyses, which is not contradictory because the latter tries to explain 100% of the variance observed.

Viral communities here studied showed a high number of exclusive vOTUs (endemism) per viral cluster or even sample (Supplementary Figure S4) and only one vOTU was found in two samples that were not part of the same cluster (Brandvlei and Porcelana at 50 ºC). Endemisms of specific viral groups (Bautista et al., 2017) and also viral communities (Bolduc et al., 2015) has been previously suggested to occur in several YNP hot springs (Bautista et al., 2017; Bolduc et al., 2015). Bautista and colleagues (2017) showed that Sulfolobus rudiviruses have a biogeographic distribution inside the YNP, while Bolduc and colleagues (2015) showed that ~62 % of the viral groups from Crater Hill Geyser and Nymph Lake are common in both samples.

This high endemism contrast with the low endemic vOTUs (15%) pattern found at global scale in the Tara Ocean marine environments dataset (Brum et al., 2015), but it is consistent with more discontinuous environment studies such as those comparing cyanophages populations from the marine coast and off shore (Gregory et al., 2016).

Together, these vOTUs results found here showed that hot springs viral communities were structured across different samples by endemic viruses at different levels (hot spring and clusters) but also by different abundances of common vOTUs. Our analyses suggest that geographic location (latitude) and temperature influences the viral populations and consequently the community structure.

Unprecedented modularity in hot springs viral network.

Monopartite protein shared networks implemented on vContact (Bolduc et al., 2017) are able to group together different viral genomes (vOTUs) into Viral Clusters (VCs), which can be considered a genus (80% precision) or a family (90% precision).

Our hot spring viral network showed an extremely high modularity with 233 connected components from 784 genomes, compared to the previous network study that analysed 1964 RefSeq genomes of

archaeal and bacterial viruses resulting in only 46 modules (Bolduc et al., 2017). In the same study, the most connected component from RefSeq network includes 1891 genomes of the order Caudovirales (Bolduc et al., 2017), while the two most connected components in our hot spring network include only 20 unclassified genomes. However, a close analysis of only the 73 archaeal viruses on Bolduc and colleagues (2017) work, show that they are divided in 8 connected components, which implies that 3.7% of all viruses occupy 17.4% of all connected components (Bolduc et al., 2017). The latter reveals a common pattern with high modularity as the one observed in our hot spring network. In fact that is not surprising, since most of genomes from archaeal viruses come from thermophilic or hypersaline systems, suggesting a high degree of modularity in viral communities from extreme ecosystems. These high modularity means that there is little or no genetic exchange between viral genomes of different genera, or in empirical words that they do not share enough proteins to establish a statistically significant relationship between their genomes. Two plausible explanations arise here for this phenomenon. First, protein based analyses reveal deep and ancient evolutionary processes, so the different lineages that we see today diverged long time ago from a common ancestor(Gregory et al., 2016). Second, in hot springs exist a tight co-evolution between viruses and their hosts, that acts as a barrier to gene flow, which in viruses is strictly restricted to episodes when two viruses co-infect the same host, requiring spatial proximity and also shared a host range (Gregory et al., 2016). The last scenario was recently proposed for cyanophages populations that infects the filamentous cyanobacteria *Fischerella* in Porcelana hot spring (Patagonia, Chile), where susceptible and resistant host populations across a 40 meters thermal gradient select specific cyanophages populations through CRISPR-Cas immunity (Guajardo-Leiva et al., 2018).

Moreover, host assignment to VCs (Figure 6) showed that some specific hosts have more than one viral genus associated and also that each VC is composed by a large number of genomes (vOTUs), which in fact represent different viral populations. The viral richness found, associated with each host, shows the existence of specific host and virus pairs capable of maintaining a high local richness. Specially interesting are the viral populations that infect common members of phototropic microbial mats such as *Fischerella, Cloroflexus* and *Meiothermus*, that are very low represented even in the large environmental databases such as IMG/VR.

827 Another interesting information obtained from our network analysis is that the VCs were composed
828 mainly of genomes of a single thermal system, genomes of nearby thermal systems (viral community
829 clusters) or only by IMG/VR genomes. Exceptionally, there were some VCs that combined genomes of
830 Brandvlei (South Africa) with genomes of Porcelana and Cahuelmó (Chile), and finally VCs with IMG/
831 VR genomes mixed with those of environmental samples. This demonstrates the existence of a
832 biogeographic effect in the distribution of the different VCs, with a few cosmopolitan genus but most
833 of them unique for each locality.
834
835 The relationship between Porcelana and Brandvlei viral genomes was also evidenced by the fact that
836 this samples shared some vPCs and at least one vOTU, which shows a potential genetic flow between
837 these two hot springs separated by 7672 km and the Atlantic Ocean between them. This suggests that
838 there is transport of viral particles, possibly by winds from Africa to South America (Griffin, 2007) and
839 that this transport has been maintained over time. The latter, together with the existence of Porcelana
840 hosts susceptible to migrant and local viruses could has maintained the cohesion between the viral
841 populations of these two hot springs.
842 These network analyzes complement and reinforce our observations on the patterns of viral community
843 structure in hot springs, and offers a framework to explore different evolutionary and population
844 genetics theories in viruses.
845
846 From a viral perspective, hot springs are constituted as natural laboratories that offer unique
847 characteristics to observe interesting evolutionary and co-evolutionary processes at different temporal
848 scales, as well as unexplored sources of proteins and enzymes whose future importance may not be
849 suspected.

850

851 **CONFLICT OF INTEREST**

852 The authors declare that the research was conducted in the absence of any commercial or financial
853 relationships that could be considered as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

SGL and BD conceived and designed the experiments. SGL performed the experiments. SGL, TAS, OS, CPA, YG, FR and BD analysed the data. SGL, and BD wrote the paper. CPA, YG and FR contributed actively with the interpretation of the results and improved the final version of the manuscript. All authors discussed the results and contributed to the final manuscript.

FUNDING

This work was financially supported by PhD scholarships CONICYT N° 21130667, 21172022 and CONICYT grant FONDECYT N°1150171.

ACKNOWLEDGMENTS

We are grateful to Huinay Scientific Field Station for making our work in the Porcelana hot spring possible.

REFERENCES

Alcamán-Arias, M. E., Pedrós-Alió, C., Tamames, J., Fernández, C., Pérez-Pantoja, D., Vásquez, M., et al. (2018). Diurnal changes in active carbon and nitrogen pathways along the temperature gradient in porcelana hot spring microbial mat. *Front. Microbiol.* 9, 1–17. doi:10.3389/fmicb.2018.02353.

Alcorta, J., Espinoza, S., Viver, T., Alcamán-Arias, M. E., Trefault, N., Rosselló-Móra, R., et al. (2018). Temperature modulates Fischerella thermalis ecotypes in Porcelana Hot Spring. *Syst. Appl. Microbiol.* 41, 531–543. doi:10.1016/j.syapm.2018.05.006.

Andam, C. P., Doroghazi, J. R., Campbell, A. N., Kelly, P. J., Choudoir, M. J., and Buckley, D. H. (2016). A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus Streptomyces. *MBio* 7, 1–9. doi:10.1128/mBio.02200-15.

Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., Breitbart, M., Salamon, P., et al. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 6, 1–9. doi:10.1186/1471-2105-6-41.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19, 455–477. doi:10.1089/cmb.2012.0021.

Bautista, M. A., Black, J. A., Youngblut, N. D., and Whitaker, R. J. (2017). Differentiation and structure in Sulfolobus islandicus rod-shaped virus populations. *Viruses* 9, 1–15. doi:10.3390/v9050120.

Bolduc, B., Jang, H. Bin, Doulcier, G., You, Z.-Q., Roux, S., and Sullivan, M. B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* 5, e3243. doi:10.7717/peerj.3243.

Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F., and Young, M. (2012). Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-Dominated Yellowstone Hot Springs. *J. Virol.* 86, 5562–5573. doi:10.1128/JVI.07196-11.

Bolduc, B., Wirth, J. F., Mazurie, A., and Young, M. J. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. *ISME J.* 9, 2162–2177. doi:10.1038/ismej.2015.28.

Breitbart, M., Wegley, L., Leeds, S., Rohwer, F., and Schoenfeld, T. (2004). Phage Community Dynamics in Hot Springs These include : Phage Community Dynamics in Hot Springs. *Appl. Environ. Microbiol.* 70, 1633–1640. doi:10.1128/AEM.70.3.1633.

Brock, T. D. (1973). Lower pH Limit for the Existence of Blue-Green Algae: Evolutionary and Ecological Implications. *Science (80-.).* 179, 480–483. doi:10.1126/science.179.4072.480.

Brum, J. R., Sullivan, M. B., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., et al. (2015). Patterns and ecological drivers of ocean viral communities. *Science (80-.).* 348, 1261498-1–11. doi:10.1126/science.1261498.

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics* 10, 1–9. doi:10.1186/1471-2105-10-421.

Davison, M., Treangen, T. J., Koren, S., Pop, M., and Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS One* 11, 1–23. doi:10.1371/journal.pone.0160574.

Diemer, G. S., and Stedman, K. M. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol. Direct* 7, 13. doi:10.1186/1745-6150-7-13.

Duffy, S., Burch, C. L., and Turner, P. E. (2007). Evolution of host specificity drives reproductive isolation among RNA viruses. *Evolution (N. Y).* 61, 2614–2622. doi:10.1111/j.1558-5646.2007.00226.x.

Duhaime, M. B., Solonenko, N., Roux, S., Verberkmoes, N. C., Wichels, A., and Sullivan, M. B. (2017). Comparative omics and trait analyses of marine Pseudoalteromonas phages advance the phage OTU concept. *Front. Microbiol.* 8, 1–16. doi:10.3389/fmicb.2017.01241.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. in genome informatics 2009, 205–211. doi:10.1142/9781848165632_0019.

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510. doi:10.1038/nrmicro1163.

Eric Wommack, K., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., et al. (2012). VIROME: A standard operating procedure for analysis of viral metagenome sequences. *Stand. Genomic Sci.* 6, 427–439. doi:10.4056/sigs.2945050.

Fuhrman, J. A., Brown, M. V., Green, J. L., Schwalbach, M. S., Brown, J. H., Steele, J. A., et al. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proc. Natl. Acad. Sci.* 105, 7774–7778. doi:10.1073/pnas.0803070105.

Gregory, A. C., Solonenko, S. A., Ignacio-Espinoza, J. C., LaButti, K., Copeland, A., Sudek, S., et al. (2016). Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 17, 930. doi:10.1186/s12864-016-3286-x.

Griffin, D. W. (2007). Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. *Clin. Microbiol. Rev.* 20, 459–477. doi:10.1128/CMR.00039-06.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, 52–57. doi:10.1093/nar/gkm360.

Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F., and Díez, B. (2018). Active crossfire between cyanobacteria and cyanophages in phototrophic mat communities within hot springs. *Front. Microbiol.* 9. doi:10.3389/fmicb.2018.02039.

Gudbergsdóttir, S. R., Menzel, P., Krogh, A., Young, M., and Peng, X. (2016). Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. *Environ. Microbiol.* 18, 863–874. doi:10.1111/1462-2920.13079.

Held, N. L., and Whitaker, R. J. (2009). Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. *Environ. Microbiol.* 11, 457–466. doi:10.1111/j.1462-2920.2008.01784.x.

Herbold, C. W., Lee, C. K., McDonald, I. R., and Cary, S. C. (2014). Evidence of global-scale aeolian dispersal and endemism in isolated geothermal microbial communities of Antarctica. *Nat. Commun.* 5, 1–10. doi:10.1038/ncomms4875.

Hildenbrand, G., Bosiek, K., Dreessen, C., Froß, P., Sievers, A., Hausmann, M., et al. (2017). K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes (Basel).* 8, 122. doi:10.3390/genes8040122.

Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. *Am. Nat.* 163, 192–211. doi:10.1086/381004.

Hurwitz, B. L., Westveld, A. H., Brum, J. R., and Sullivan, M. B. (2014). Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc. Natl. Acad. Sci.* 111, 10714–10719. doi:10.1073/pnas.1319778111.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* 12, 1–12. doi:10.1371/journal.pcbi.1004957.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11. doi:10.1186/1471-2105-11-119.

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., and Rusch, D. B. (2013). The YNP metagenome project: Environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. *Front. Microbiol.* 4, 1–15. doi:10.3389/fmicb.2013.00067.

Inskeep, W. P., Rusch, D. B., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Richardson, T. H., et al. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. *PLoS One* 5. doi:10.1371/journal.pone.0009773.

Iranzo, J., Krupovic, M., and Koonin, E. V (2016). The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. 7, 1–21. doi:10.1128/mBio.00978-16.Editor.

Jiang, B., Song, K., Ren, J., Deng, M., Sun, F., and Zhang, X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 13. doi:10.1186/1471-2164-13-730.

Jun, S.-R., Sims, G. E., Wu, G. A., and Kim, S.-H. (2009). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci.* 107, 133–138. doi:10.1073/pnas.0913033107.

Klatt, C. G., Inskeep, W. P., Herrgard, M. J., Jay, Z. J., Rusch, D. B., Tringe, S. G., et al. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. *Front. Microbiol.* 4, 1–23. doi:10.3389/fmicb.2013.00106.

Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME J.* 11, 2399–2406. doi:10.1038/ismej.2017.113.

Koonin, E. V, Senkevich, T. G., and Dolja, V. V (2006). The ancient Virus World and evolution of cells. *Biol. Direct* 1, 1–27. doi:https://doi.org/10.1186/1745-6150-1-29.

983 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al. (2004).
984 Versatile and open software for comparing large genomes. *Genome Biol.* 5, R12. doi:10.1186/gb-
985 2004-5-2-r12.

986 Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9,
987 357–9. doi:10.1038/nmeth.1923.

988 Li, D., Luo, R., Liu, C. M., Leung, C. M., Ting, H. F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A
989 fast and scalable metagenome assembler driven by advanced methodologies and community
990 practices. *Methods* 102, 3–11. doi:10.1016/j.ymeth.2016.02.020.

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

993 Mackenzie, R., Pedrós-Alió, C., and Díez, B. (2013). Bacterial composition of microbial mats in hot
994 springs in Northern Patagonia: Variations with seasons and temperature. *Extremophiles* 17, 123–
995 136. doi:10.1007/s00792-012-0499-z.

996 Maltez Thomas, A., Prata Lima, F., Maria Silva Moura, L., Maria da Silva, A., Dias-Neto, E., and
997 Setubal, J. C. (2018). Comparative metagenomics. *Methods Mol. Biol.* 1704, 243–260.
998 doi:10.1007/978-1-4939-7463-4_8.

999 Markine-Goriaynoff, N., Gillet, L., Van Etten, J. L., Korres, H., Verma, N., and Vanderplasschen, A.
1000 (2004). Glycosyltransferases encoded by viruses. *J. Gen. Virol.* 85, 2741–2754.
1001 doi:10.1099/vir.0.80320-0.

1002 Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads.
1003 *EMBnet.journal* 17, 10. doi:10.14806/ej.17.1.200.

1004 Mavrich, T. N., and Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and
1005 genome. *Nat. Microbiol.* 2, 1–9. doi:10.1038/nmicrobiol.2017.112.

1006 Menzel, P., Gudbergsdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., et al. (2015).
1007 Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs. *Microb.*
1008 *Ecol.* 70, 411–424. doi:10.1007/s00248-015-0576-9.

1009 Munson-Mcgee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J., Weitz, J. S., et al.
1010 (2018). A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in
1011 extreme environments. *ISME J.* 12, 1706–1714. doi:10.1038/s41396-018-0071-7.

1012 Oksanen J, Blanchet G, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R, Simpson
1013 G, Solymos P, Stevens H, Szoecs E and Wagner H. (2019). vegan: Community Ecology Package.
1014 R package version 2.5-5. https://CRAN.R-project.org/package=vegan

O'Malley, M. A. (2008). "Everything is everywhere: but the environment selects": ubiquitous distribution and ecological determinism in microbial biogeography. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* 39, 314–325. doi:10.1016/j.shpsc.2008.06.005.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17, 1–14. doi:10.1186/s13059-016-0997-x.

Paez-Espino, D., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., Szeto, E., et al. (2017). IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* 45, D457–D465. doi:10.1093/nar/gkw1030.

Paez-espino, D., Roux, S., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2019). IMG / VR v . 2 . 0 : an integrated data management and analysis system for cultivated and environmental viral genomes. 47, 678–686. doi:10.1093/nar/gky1127.

Papke, R. T., Ramsing, N. B., Bateson, M. M., and Ward, D. M. (2003). Geographical isolation in hot spring cyanobacteria. *Environ. Microbiol.* 5, 650–659. doi:10.1046/j.1462-2920.2003.00460.x.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114.

Power, J. F., Carere, C. R., Lee, C. K., Wakerley, G. L. J., Evans, D. W., Button, M., et al. (2018). Microbial biogeography of 925 geothermal springs in New Zealand. *Nat. Commun.* 9. doi:10.1038/s41467-018-05020-y.

Pride, D. T., and Schoenfeld, T. (2008). Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics* 9, 1–16. doi:10.1186/1471-2164-9-420.

Richter, M., and Rossello-Mora, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci.* 106, 19126–19131. doi:10.1073/pnas.0906412106.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A., and Sullivan, M. B. (2017). Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* 5, e3817. doi:10.7717/peerj.3817.

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 7. doi:10.1371/journal.pone.0033641.

Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. (2018). Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. *Nat. Ecol. Evol.* 2, 510–519. doi:10.1038/s41559-017-0435-9.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 2011, 27:863-864. *Bioinformatics* 27, 863–864. doi:10.1093/bioinformatics/btq281.2.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., and Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. *Appl. Environ. Microbiol.* 74, 4164–4174. doi:10.1128/AEM.02598-07.

Shannon, P., Markiel, A., Owen Ozier, 2, Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 2498–2504. doi:10.1101/gr.1239303.metabolite.

Sharp, C. E., Brady, A. L., Sharp, G. H., Grasby, S. E., Stott, M. B., and Dunfield, P. F. (2014). Humboldt's spa: Microbial diversity is controlled by temperature in geothermal environments. *ISME J.* 8, 1166–1174. doi:10.1038/ismej.2013.237.

Sievers, A., Wenz, F., Hausmann, M., and Hildenbrand, G. (2018). Conservation of k-mer Composition and Correlation Contribution between Introns and Intergenic Regions of Animalia Genomes. *Genes (Basel).* 9, 482. doi:10.3390/genes9100482.

Snyder, J. C., Wiedenheft, B., Lavin, M., Roberto, F. F., Spuhler, J., Ortmann, A. C., et al. (2007). Virus movement maintains local virus population diversity. *Proc. Natl. Acad. Sci.* 104, 19102–19107. doi:10.1073/pnas.0709445104.

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

Whitaker, R. J. (2003). Geographic Barriers Isolate Endemic Populations of Hyperthermophilic Archaea. *Science (80-.).* 301, 976–978. doi:10.1126/science.1086909.

Wilson, M. S., Siering, P. L., White, C. L., Hauser, M. E., and Bartles, A. N. (2008). Novel archaea and bacteria dominate stable microbial communities in North America's largest hot spring. *Microb. Ecol.* 56, 292–305. doi:10.1007/s00248-007-9347-6.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi:10.1093/bioinformatics/btv638.

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The Sorcerer II global ocean sampling expedition: Expanding the universe of protein families. *PLoS Biol.* 5, 0432–0466. doi:10.1371/journal.pbio.0050016.

Zablocki, O., van Zyl, L. J., Kirby, B., and Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. *Viruses* 9. doi:10.3390/v9110348.

Zablocki, O., van Zyl, L., and Trindade, M. (2018). Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages. *Extremophiles* 22, 827–837. doi:10.1007/s00792-018-1052-5.

Zaslavsky, L., Ciufo, S., Fedorov, B., and Tatusova, T. (2016). Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *BMC Bioinformatics* 17. doi:10.1186/s12859-016-1112-8.

Data Availability Statement

The datasets generated for this study can be found NCBI as follow: Access to raw data for metagenomes and metatranscriptomes is available through NCBI BioProject ID PRJNAxxxx

FIGURES

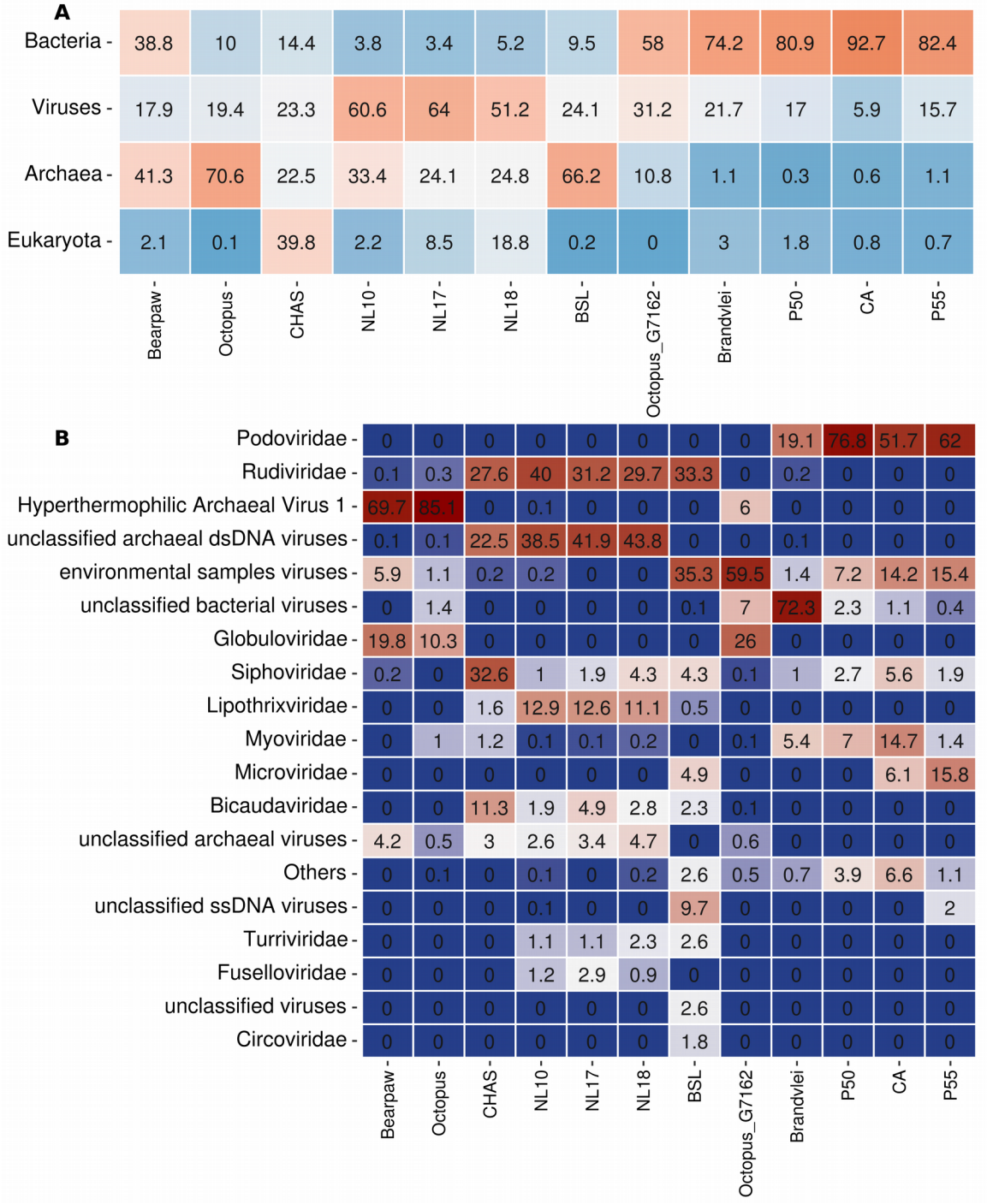


Figure 1. Relative abundances of protein sequences in viral metagenomes, classified by LCA algorithm trough local alignment to NCBI nr database. A) Domain level, and B) Family level for sequences classified as Virus in A. Sequences were normalized by protein length and library size. Families with abundances below 0.1% were summarized as others.

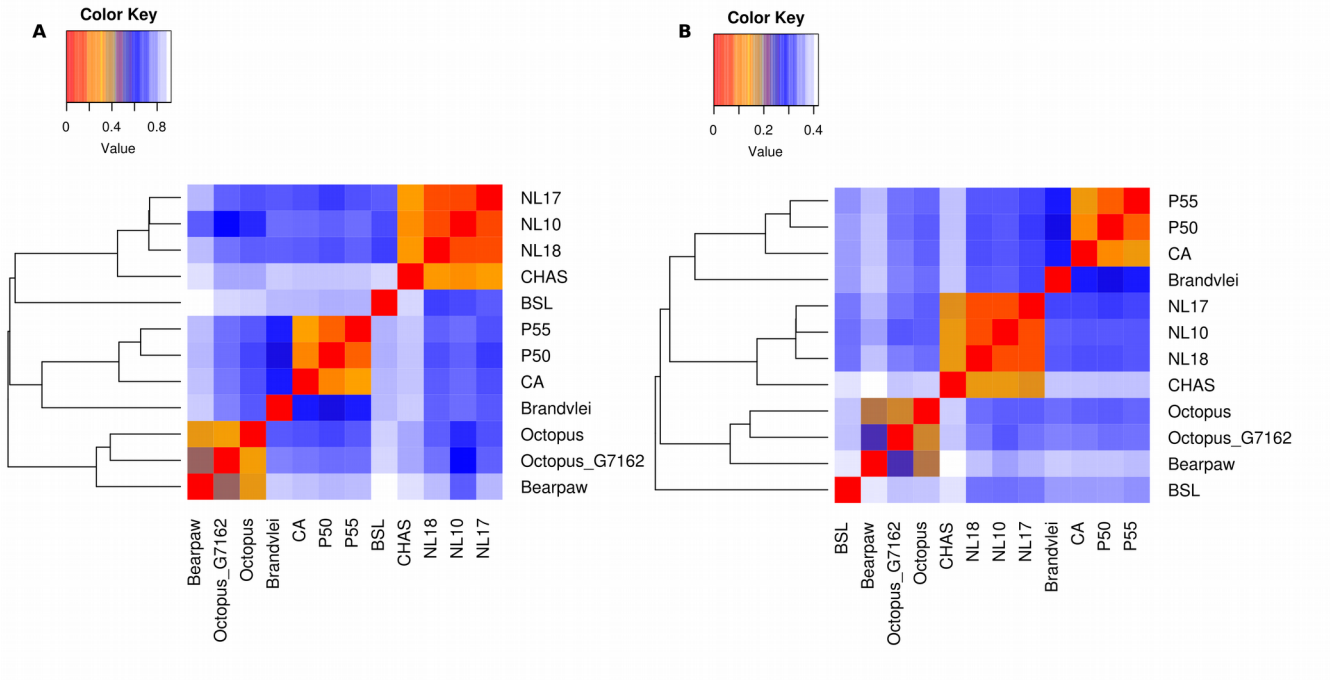


Figure 2. Hierarchical clustering of hot springs (A) proteins and (B) genes. Dendogram was constructed based on MASH distance matrix of 12 samples and 161391 proteins or genes. Heatmap was colored according to MASH dissimilarity index.

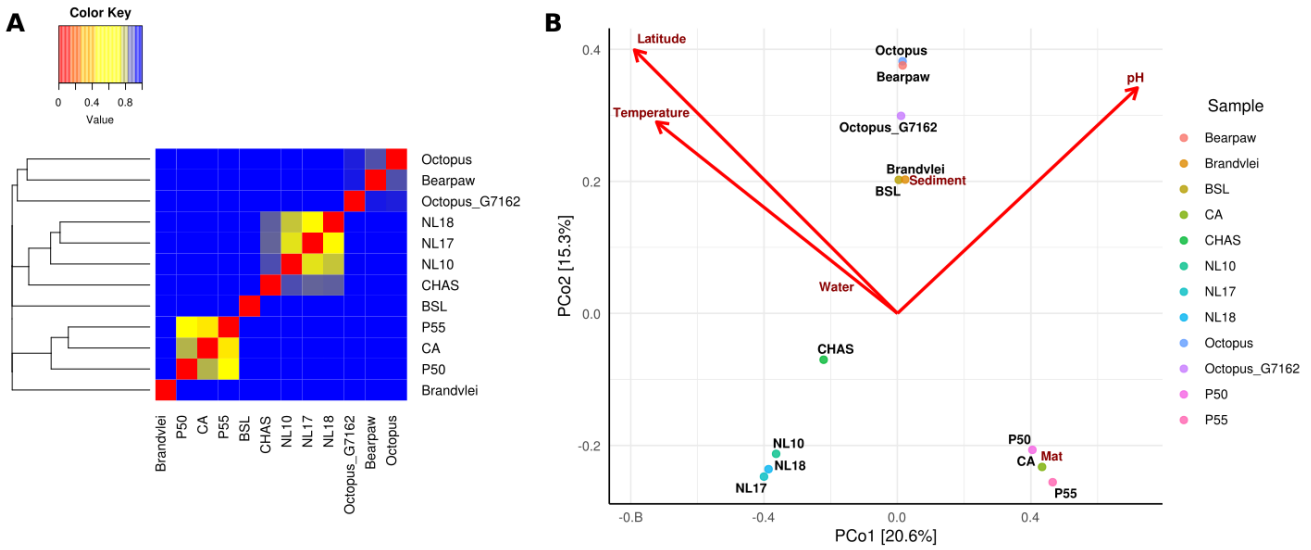


Figure 3. Viral community differentiation measured through vPCs based on to Bray-Curtis distance matrix of 12 samples and 22667 vPCs**. A)**Hierarchical clustering of hot sprigs vPCs. Dendrogram was constructed based on to Bray-Curtis distance matrix and heatmap is colored according to Bray-Curtis dissimilarity. **B)** Principal coordinate analyses of hot springs vPCs. PCoA was constructed based on to Bray-Curtis distance matrix. No initial data transformation has been applied. The relative contribution (eigenvalue) of each axis to the total inertia in the data is indicated as percent at the axis titles. Vectors of environmental variables ($P < 0.05$) were fitted onto ordination space.

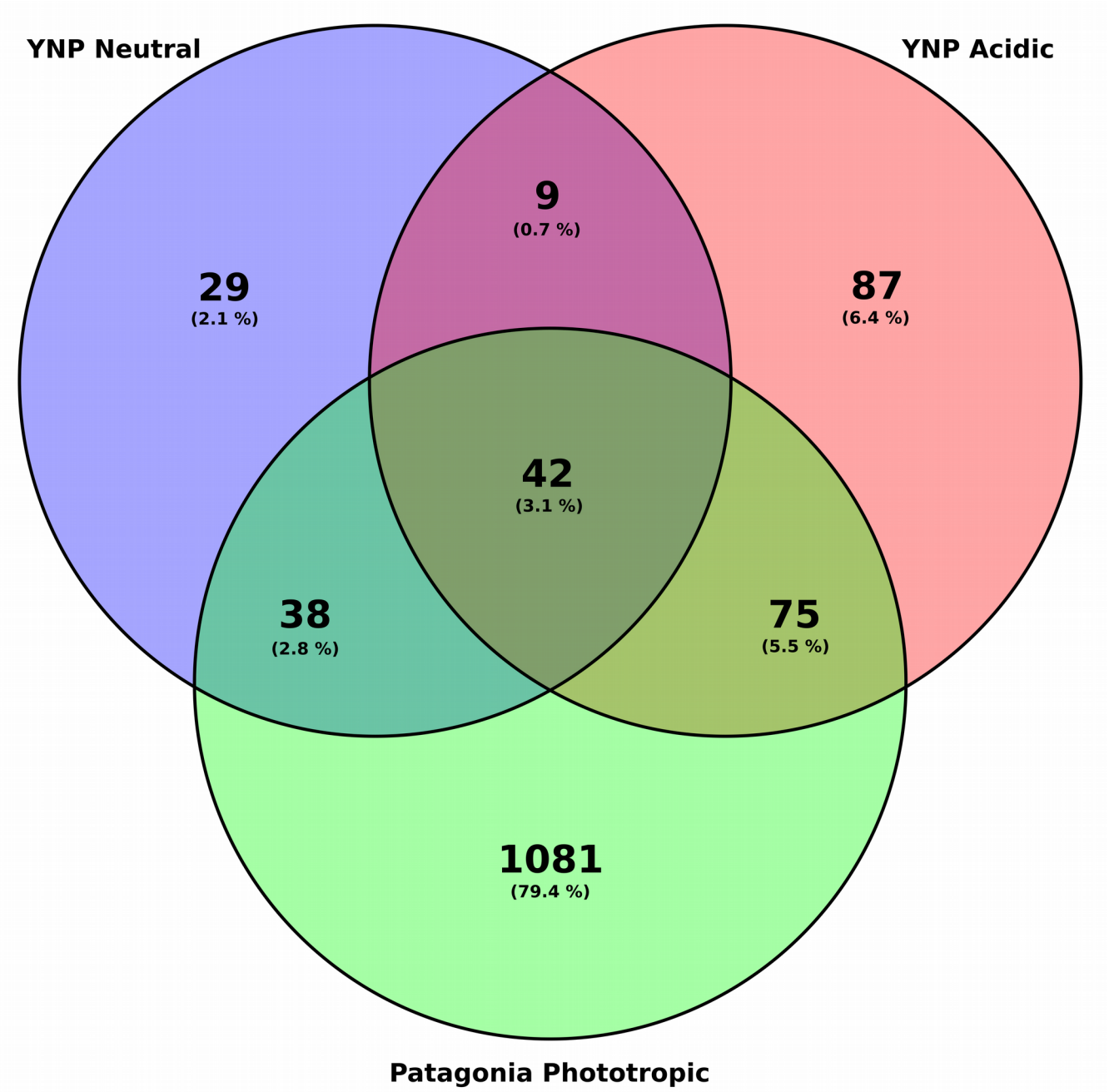


Figure 4. Venn diagram of Pfam classification of hot spring vPCs, that correspond to 1361 protein families. Each circle represent a hot spring cluster: YNP Acidic (NL10, NL17, NL18 and CHAS), YNP Neutral (Bear paw, Octopus, and Octopus G7162) and Patagonia Phototrophic (P50, P55, CA).

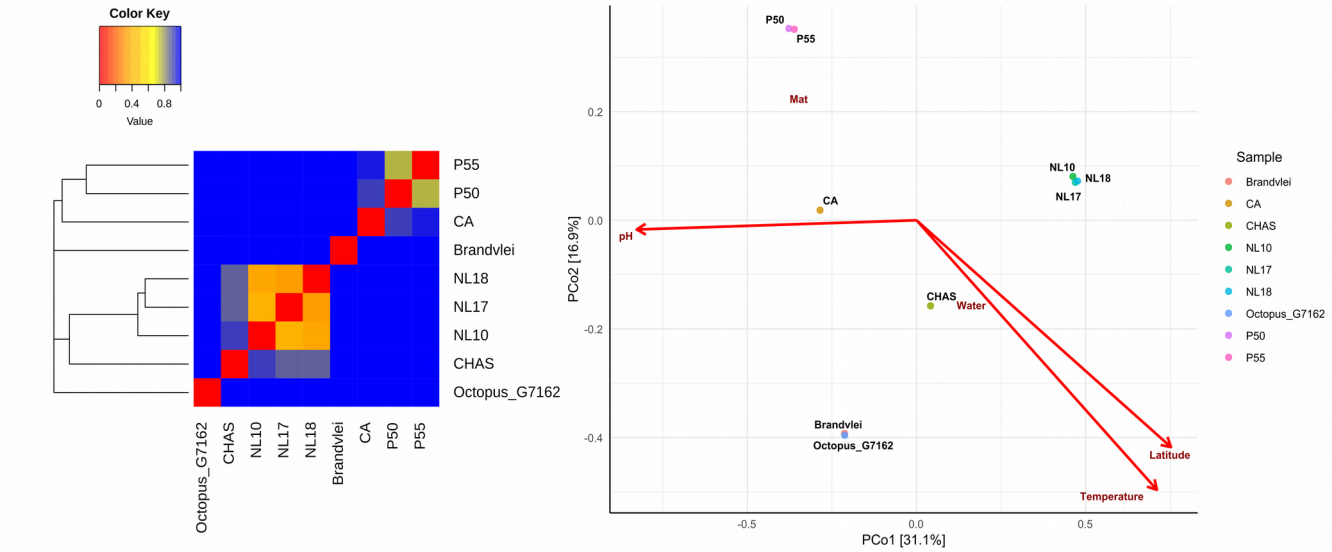


Figure 5. Viral community structure measured through vOTUs based on to Bray-Curtis distance matrix of 9 samples and 948 vOTUs**. A)**Hierarchical clustering of hot sprigs vOTUs. Dendrogram was constructed based on to Bray-Curtis distance matrix and heatmap is colored according to Bray-Curtis dissimilarity. **B)** Principal coordinate analyses of hot springs vOTUs. PCoA was constructed based on to Bray-Curtis distance matrix. No initial data transformation has been applied. The relative contribution (eigenvalue) of each axis to the total inertia in the data is indicated in percent at the axis titles. Vectors of environmental variables ($P < 0.05$) were fitted onto ordination space.

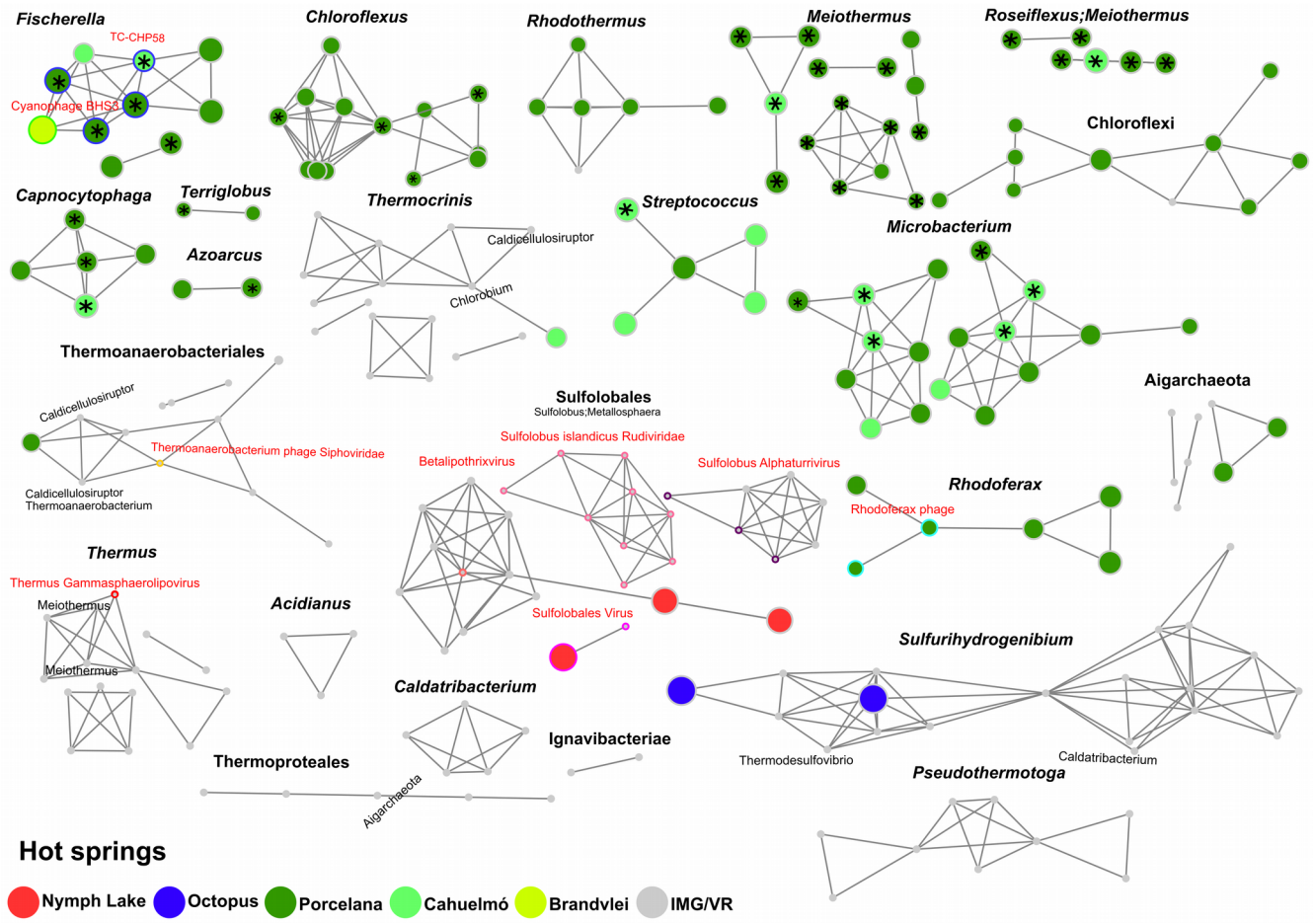


Figure 6. Subset of a protein-sharing network for 784 (genomes) vOTUs from 9 hot springs and IMG/VR contigs. Each node represents a vOTU, the color of each node represent the hot spring from it was recovered (as indicated in the legend) and node size represent the logarithmic transformation of the genome relative abundance across all samples. Edges between nodes indicate a statistically significant relationship between the protein profiles of their viral genomes. Modules within the network are composed of groups of similar sequences using the MCL algorithm with an inflation value of 2. Names in bold black letters represent the predicted host by the use of CRISPR spacers match and asterisk symbols indicate vOTUs with CRISPR spacer(s) match. Names in red letters and border color of nodes, represent different best BLASTN hit in viral RefSeq database.

TABLES

Table 1. Hot springs environmental factors, sample source and geographic location.

Sample	Temperature ºC	pH	Source	Latitude	Longitude
BSL	52	2	Sediment	40.436	-121.398
CHAS	82	2	Water	44.651	-110.485
NL10	86	3.56	Water	44.754	-110.724
NL17	80	2.25	Water	44.7521	-110.729
NL18	81	2.25	Water	44.7521	-110.729
Octopus	93	8.14	Water	44.534	-110.798
Octopus G7162	84	8.1	Water	44.534	-110.798
Bear paw	74	7.34	Water	44.556	-110.835
P50	50	7.24	Mat	-42.462	-72.467
P55	55	6.67	Mat	-42.462	-72.467
CA	51	9.37	Mat	-42.26	-72.386
Brandvlei	60	5.7	Water	-33.732	19.413

Table 2. Permanova analyses (Type III, marginal test) of hot springs properties over vPCs beta-diversity (Bray-Curtis).

Properties	Df	Sum Of Sqs	R^2	F	Pr(>F)
Temperature	1	0.3205	0.06227	0.712	0.9023
pH	1	0.6362	0.12362	1.4134	0.0735
Source	2	0.7447	0.14469	0.8271	0.8715
Latitude	1	0.3333	0.06475	0.7403	0.8772
Residual	6	2.7008	0.52479		
Total	11	5.1465	1		

Table 3. Permanova analyses (Type III, marginal test) of hot springs properties over vOTUs beta-diversity (Bray-Curtis).

Properties	Df	Sum Of Sqs	R^2	F	Pr(>F)
Temperature	1	0.5682	0.15946	1.1761	0.2904
pH	1	0.3519	0.09877	0.7285	0.7926
Source	1	0.4944	0.13875	1.0233	0.5146
Latitude	1	0.6209	0.17426	1.2852	0.2140
Residual	4	1.9324	0.54234		
Total	8	3.5631	1		

Supplementary Material

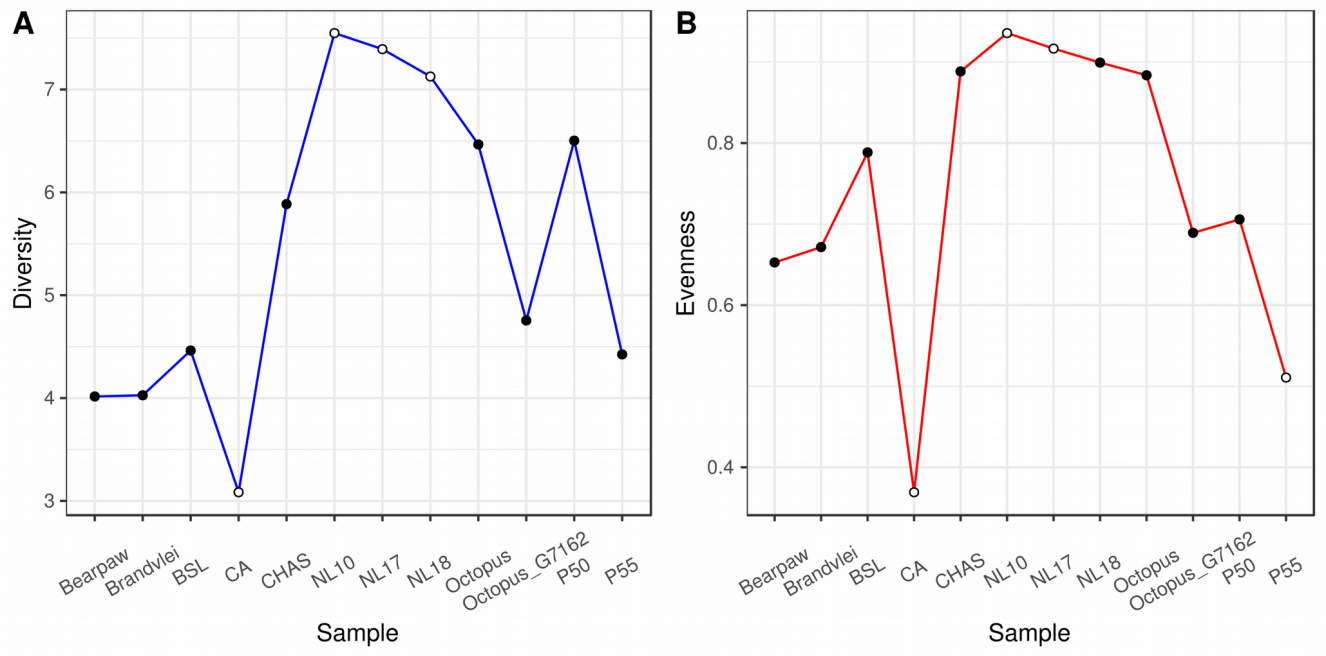
Ecological drivers modulate biogeography in thermophilic viral communities

Sergio Guajardo-Leiva, Oscar Salgado, Tomás Alarcón-Schumacher, Juris A. Grasis, Forest Rohwer and Beatriz Díez.

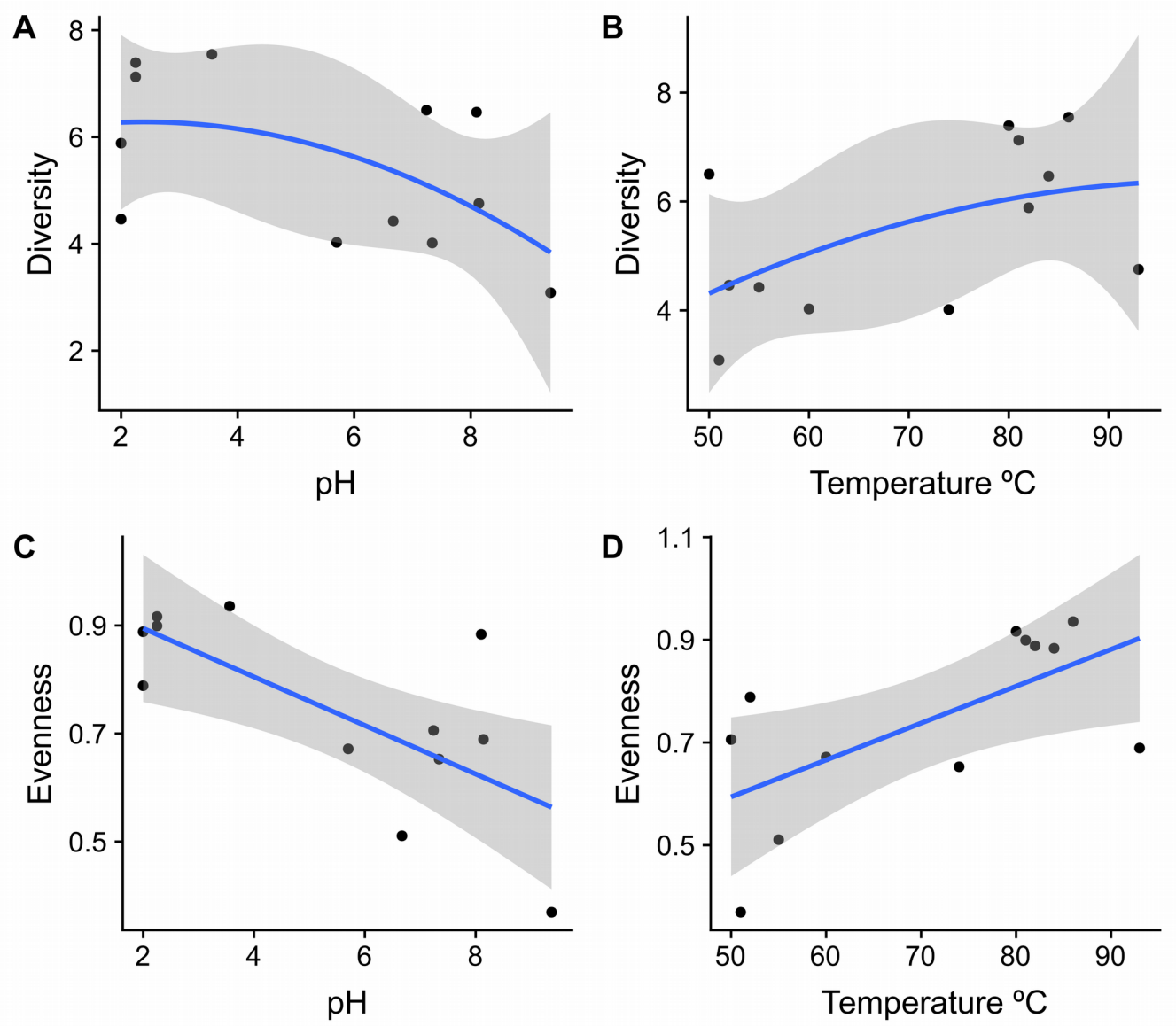
*** Correspondence:** Beatriz Díez: bdiez@bio.puc.cl

Supplementary Figures and Tables

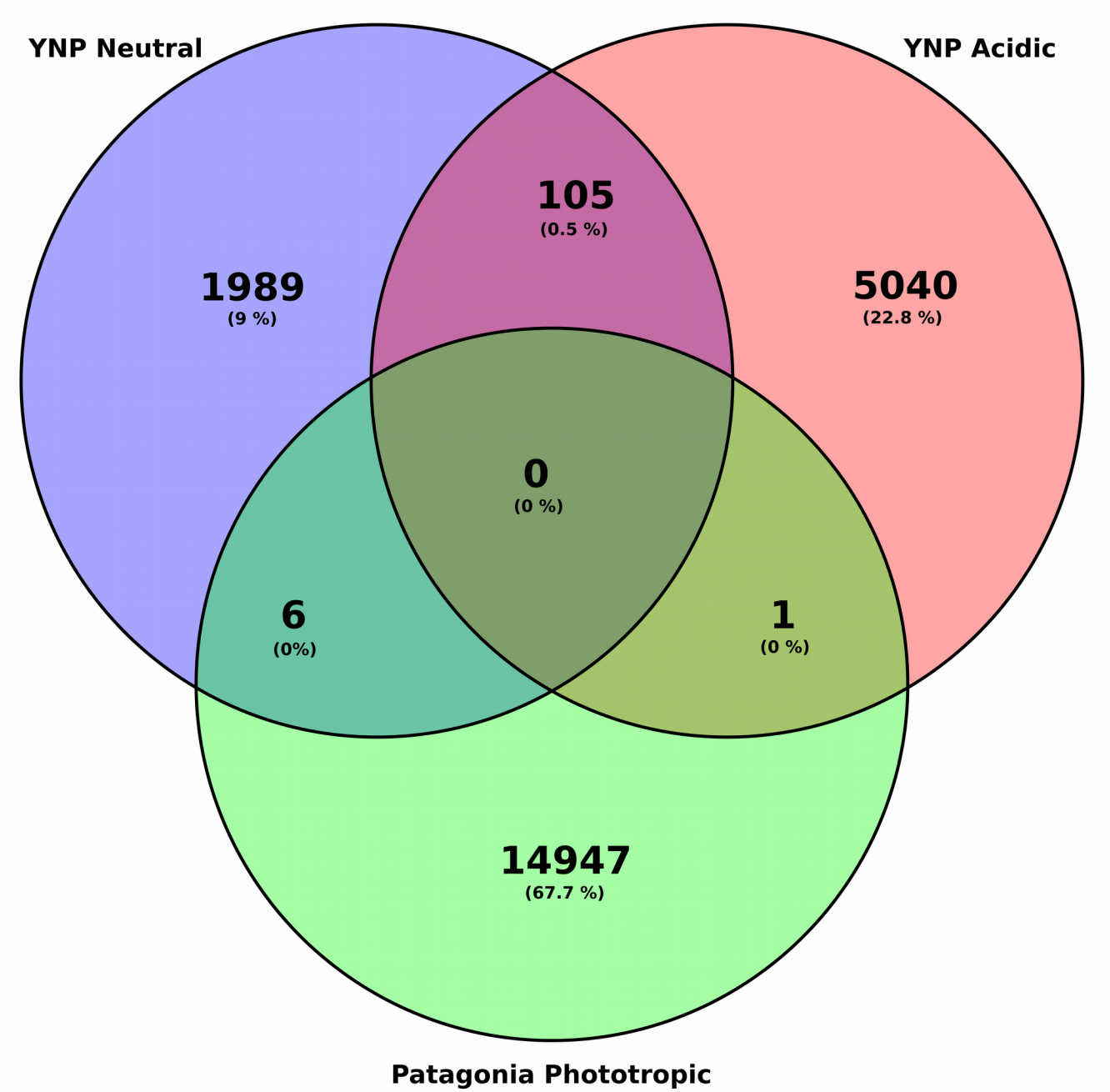
Supplementary Figures



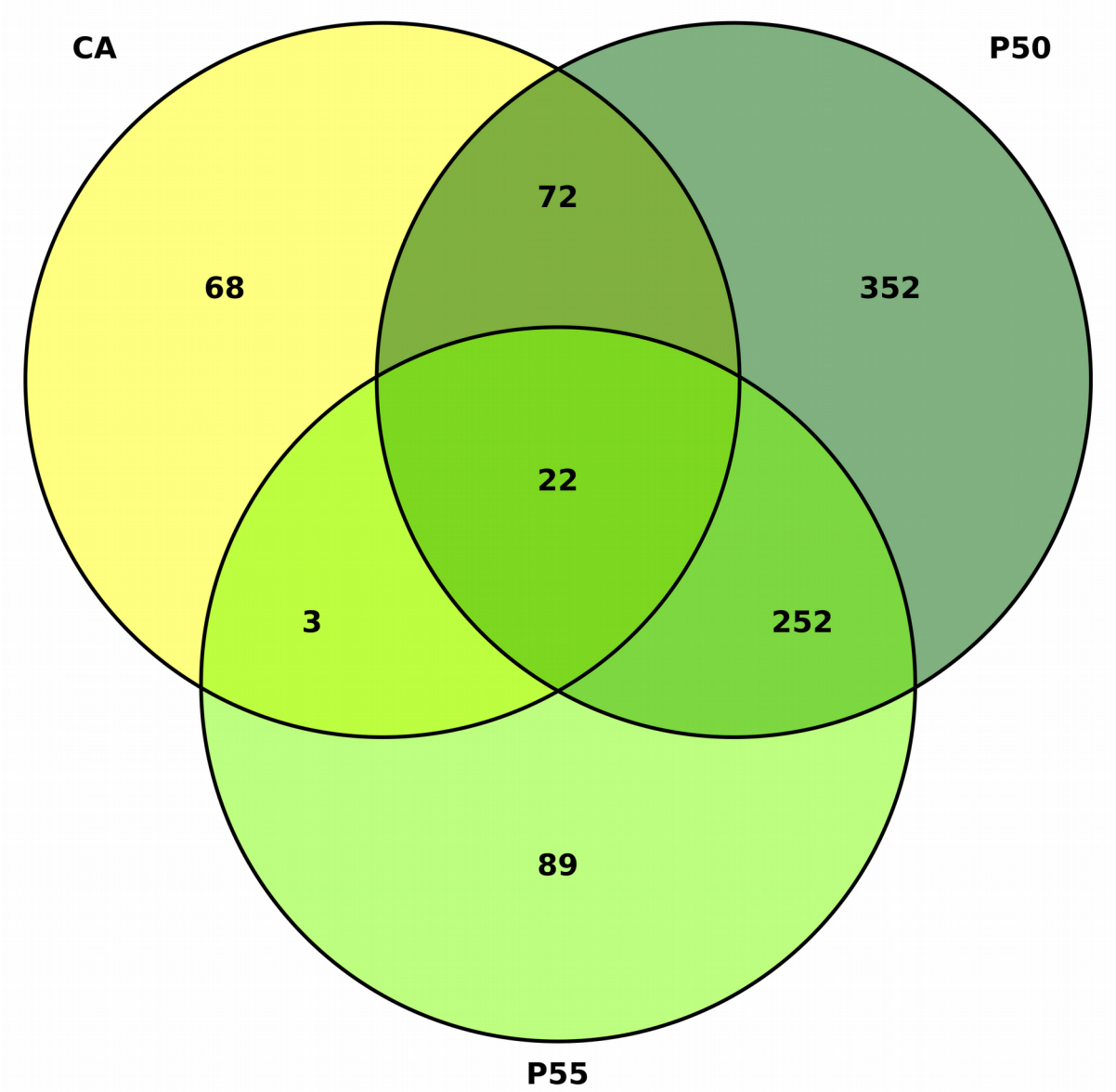
Supplementary Figure S1. Alpha diversity measurements in hot springs viral metagenomes. vPCs normalized counts for each sample were used to calculate A) Shannon's diversity and B) Pielou's evenness. Outliers corresponding to diversity and evenness values outside the average SD are represented as clear circles.



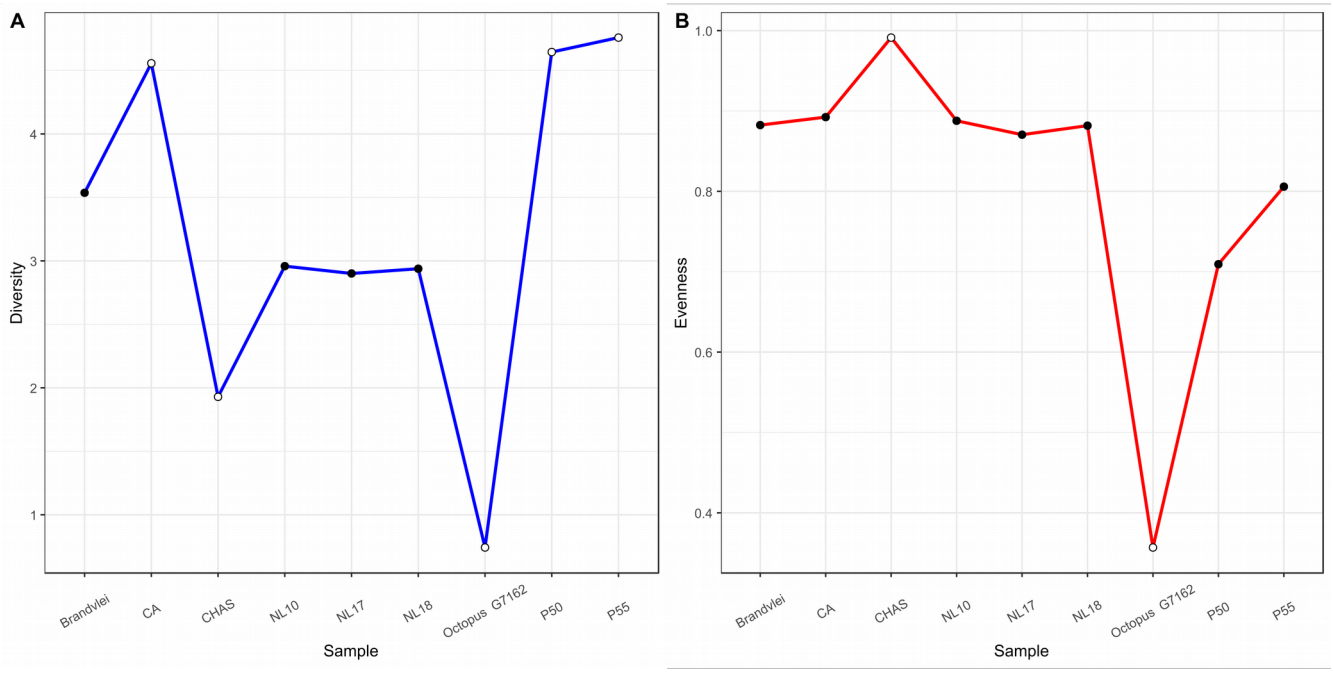
Supplementary Figure S2. The relationship between hot springs physico-chemical properties and vPCs diversity and evenness. A) Quadratic regression of vPCs diversity (Shannon's index) and hot springs pH. B) Quadratic regression of vPCs diversity (Shannon's index) and hot springs temperature. C)Linear regression of vPCs evenness (Pielou's) and hot springs pH. D) Linear regression of vPCs evenness (Pielou's) and hot springs temperature.



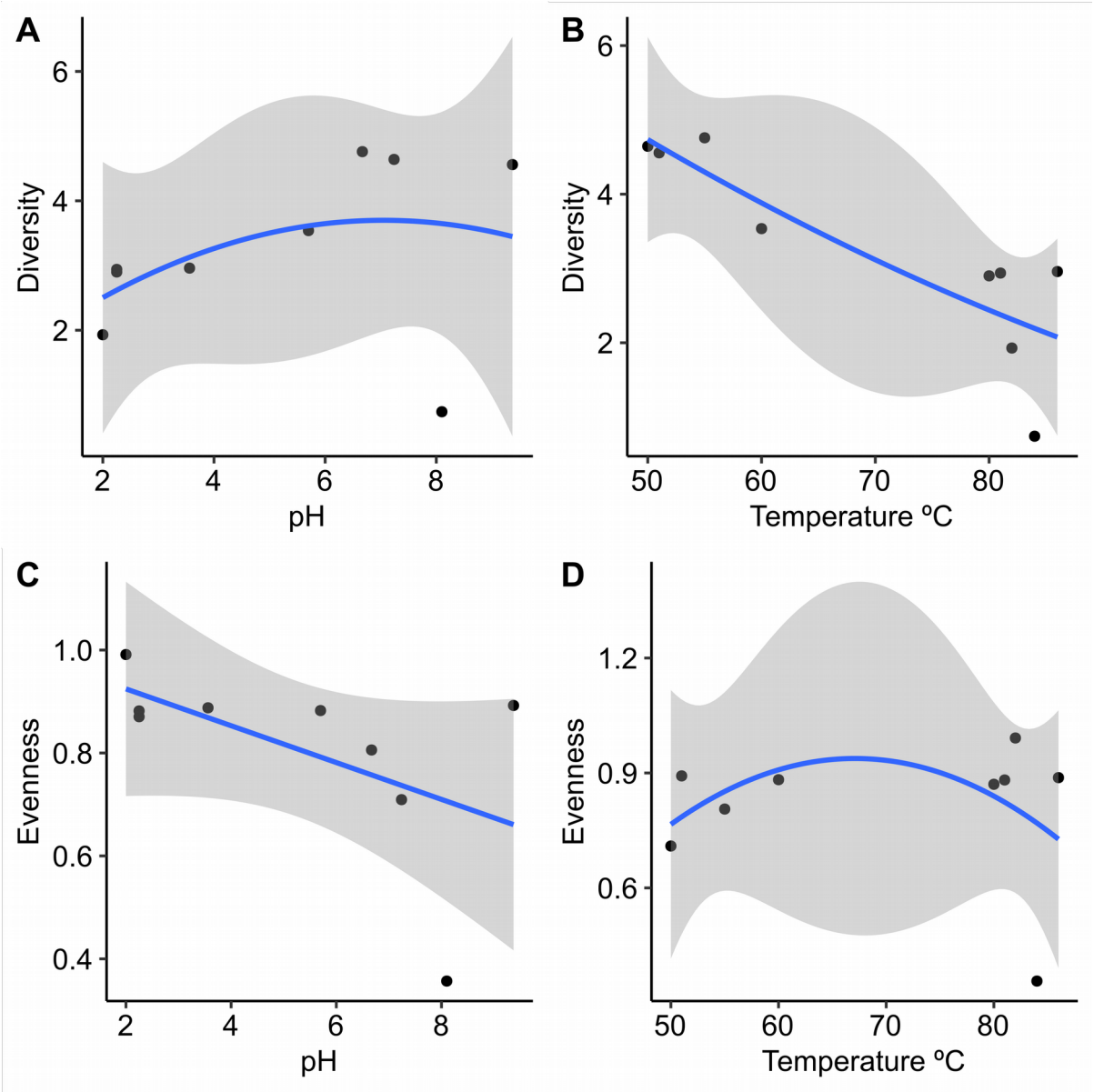
Supplementary Figure S3. Venn diagram of hot springs vPCs, that correspond to 1361 protein families. Each circle represent a hot spring cluster: YNP Acidic (NL10, NL17, NL18 and CHAS), YNP Neutral (Bear paw, Octopus, and Octopus G7162) and Patagonia Phototropic (P50, P55, CA).



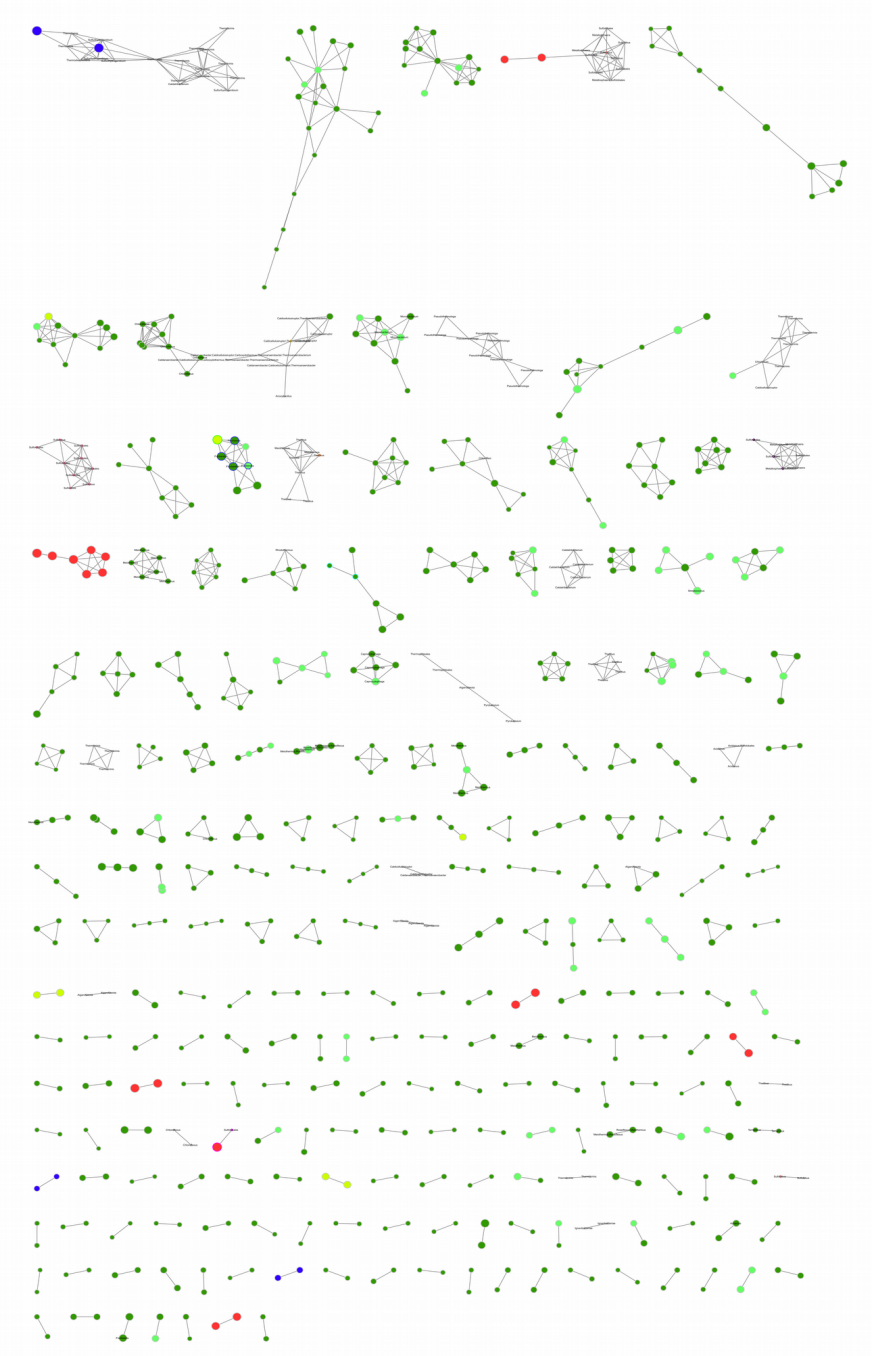
Supplementary Figure S4. Venn diagram of Patagonia Phototropic hot springs vOTUs, that correspond to 858 sequences. Each circle represent a hot spring sample inside de cluster: (CA) Cahuelmó, (P50) Porcelana 50 ºC and P55 Porcelana 55 ºC.



Supplementary Figure S5. Alpha diversity measurements in hot springs viral metagenomes. vOTUs normalized counts for each sample were used to calculate A) Shannon's diversity and B) Pielou's evenness. Outliers corresponding to diversity and evenness values outside the average SD are represented as clear circles.



Supplementary Figure S6. The relationship between hot springs physico-chemical properties and vOTUs diversity and evenness. A) Quadratic regression of vOTUs diversity (Shannon's index) and hot springs pH. B) Quadratic regression of vOTUs diversity (Shannon's index) and hot springs temperature. C)Linear regression of vOTUs evenness (Pielou's) and hot springs pH. D) Quadratic regression of vOTUs evenness (Pielou's) and hot springs temperature.



Supplementary Figure S7. Protein-sharing network for 784 (genomes) vOTUs from 9 hot springs and IMG/VR contigs. Each node represents a vOTU or IMG/VR contig and the color of each node represent the hot spring from it was recovered. Edges between nodes indicate a statistically significant relationship between the protein profiles of their viral genomes. Modules within the network are composed of groups of similar sequences using the MCL algorithm with an inflation value of 2. Names in bold black letters represent the predicted host by the use of CRISPR spacers match and asterisk symbols indicate vOTUs with CRISPR spacer(s) match.

Supplementary Tables

Supplementary Table S1. Summary information about sequencing depth, quality filtering, read mapping and assembly of hot springs viral metagenomes.

Sample	Raw sequences (M)	Quality filtered sequences (M)	Quality filtered Bases (M)	Nº of viral contigs ≥ 500pb	Nº of predicted proteins	NCBI nr aligned sequences (k)	VPCs recruitment (k)
Bear paw	0.0084	0.0084	8.27	1758	2815	6.15	5.99
Brandvlei	0.8700	0.87	199.86	7973	13993	70.60	36.84
BSL	0.3900	0.39	136.22	2735	4783	30.80	180.75
CA	2.3400	2.14	505.65	10805	20417	91.42	1030.83
CHAS	0.1900	0.19	56.63	2399	3564	31.90	39.28
NL10	0.4552	0.4552	148.22	4978	9939	110.83	184.58
NL17	0.3993	0.3993	140.63	9219	14918	129.26	183.21
NL18	0.4705	0.4705	159.14	7405	11833	130.38	189.47
Octopus	0.0220	0.022	22.56	5975	10163	19.164	20.3
Octopus G7162	0.2300	0.23	86.16	4048	7301	45.95	106.75
P50	4.1300	3.67	849.47	16894	40485	742.78	1507.26
P55	3.0100	2.77	651.65	9629	20820	311.42	1477.61

Supplementary Table S2: Hot springs viral metagenomes information.

Sample	Reference	Accession	Link
Bear paw	Schoenfeld at al ., 2008	JGI_LIB_BEARPAW_20031007	https://www.imicrobe.us/#/samples/1276
Brandvlei	Zablocki et al ., 2017.	PRJEB18453	https://www.ebi.ac.uk/ena/data/view/PRJEB18453
BSL	Diemer and Stedman., 2012.	CAM_SMPL_000802 CAM_SMPL_000830	https://www.imicrobe.us/#/samples/311 https://www.imicrobe.us/#/samples/339
CA	This study		https://www.ncbi.nlm.nih.gov/sra/
CHAS	M. Young, Montana State University. Unpublished	CAM_SMPL_000984	https://www.imicrobe.us/#/samples/387
NL10	Bolduc et al., 2012.	CAM_SMPL_000955 CAM_SMPL_000983 CAM_SMPL_001008	https://www.imicrobe.us/#/samples/358 https://www.imicrobe.us/#/samples/386 https://www.imicrobe.us/#/samples/411
NL17	Bolduc et al., 2012.	CAM_SMPL_000976 CAM_SMPL_001001	https://www.imicrobe.us/#/samples/379 https://www.imicrobe.us/#/samples/404
NL18	Bolduc et al., 2012.	CAM_SMPL_000967 CAM_SMPL_001002	https://www.imicrobe.us/#/samples/370 https://www.imicrobe.us/#/samples/405
Octopus	Schoenfeld at al ., 2008	JGI_LIB_OCTOPUS_20031004	https://www.imicrobe.us/#/samples/1277
Octopus G7162	Schoenfeld, Lucigen Corporation. Unpublished	CAM_SMPL_000836	https://www.imicrobe.us/#/samples/345
P50	This study		https://www.ncbi.nlm.nih.gov/sra/
P55	This study		https://www.ncbi.nlm.nih.gov/sra/

Supplementary Table S3. Pfam protein families (42) common in to all hot spring clusters of samples. Bold text highlights hallmark genes.

Pfam Accession	Description
PF00004.28	ATPase family associated with various cellular activities (AAA)
PF00082.21	Subtilase family
PF00085.19	Thioredoxin
PF00136.20	**DNA polymerase family B**
PF00145.16	C-5 cytosine-specific DNA methylase
PF00216.20	Bacterial DNA-binding protein
PF00239.20	Resolvase, N terminal domain
PF00271.30	Helicase conserved C-terminal domain
PF00476.19	DNA polymerase family A
PF00534.19	Glycosyl transferases group 1
PF00589.21	**Phage integrase family**
PF00692.18	dUTPase
PF01022.19	Bacterial regulatory protein, arsR family
PF01402.20	Ribbon-helix-helix protein, copG family
PF01507.18	Phosphoadenosine phosphosulfate reductase family
PF01870.17	Archaeal holliday junction resolvase (hjc)
PF01930.16	Domain of unknown function DUF83
PF01935.16	Helicase HerA, central domain
PF02086.14	DNA methyltransferase
PF02195.17	ParB-like nuclease domain
PF02481.14	DNA recombination-mediator protein A
PF02511.14	Thymidylate synthase complementing protein
PF02867.14	Ribonucleotide reductase, barrel domain
PF04434.16	SWIM zinc finger
PF04466.12	**Phage terminase large subunit**
PF04513.11	Baculovirus polyhedron envelope protein, PEP, C terminus
PF04586.16	**Caudovirus prohead serine protease**
PF04851.14	Type III restriction enzyme, res subunit
PF05707.11	Zonular occludens toxin (Zot)
PF05866.10	Endodeoxyribonuclease RusA
PF06048.10	Domain of unknown function (DUF927)
PF07728.13	AAA domain (dynein-related subfamily)
PF08423.10	Rad51
PF08960.9	Domain of unknown function (DUF1874)
PF09250.10	Bifunctional DNA primase/polymerase, N-terminal
PF10102.8	Domain of unknown function (DUF2341)
PF12705.6	PD-(D/E)XK nuclease superfamily
PF13307.5	Helicase C-terminal domain
PF13412.5	Winged helix-turn-helix DNA-binding
PF13489.5	Methyltransferase domain
PF13521.5	AAA domain
PF13692.5	Glycosyl transferases group 1

Supplementary Table S4. Relative abundance of vPCs with Pfam annotation, only Pfam families with total abundance ≥1% are show. Bold text highlights hallmark genes.

Pfam Accession	Description	% of abundance
PF06048	Domain of unknown function (DUF927)	3.9
PF08960	Domain of unknown function (DUF1874)	3.6
PF01402	Ribbon-helix-helix protein, copG family	3.4
PF00589	**Phage integrase family**	2.7
PF04466	**Phage terminase large subunit**	2.6
PF12193	Sulfolobus virus coat protein C terminal	2.4
PF08281	Sigma-70, region 4	1.9
PF00534	Glycosyl transferases group 1	1.9
PF12705	PD-(D/E)XK nuclease superfamily	1.6
PF12796	Ankyrin repeats (3 copies)	1.6
PF09250	Bifunctional DNA primase/polymerase, N-terminal	1.5
PF00004	ATPase family associated with various cellular activities (AAA)	1.5
PF17289	Terminase RNaseH-like domain	1.5
PF13521	AAA domain	1.5
PF01022	Bacterial regulatory protein, arsR family	1.2
PF00515	Tetratricopeptide repeat	1.2
PF01507	Phosphoadenosine phosphosulfate reductase family	1.2
PF02195	ParB-like nuclease domain	1.2
PF01870	Archaeal holliday junction resolvase (hjc)	1.2
PF12236	**Bacteriophage head to tail connecting protein**	1.1
PF02768	DNA polymerase III beta subunit, C-terminal domain	1.1
PF04513	**Baculovirus polyhedron envelope protein, PEP, C terminus**	1.1
PF02086	D12 class N6 adenine-specific DNA methyltransferase	1.1
PF00176	SNF2 family N-terminal domain	1.0
PF00535	Glycosyl transferase family 2	1.0
PF04851	Type III restriction enzyme, res subunit	1.0
PF13692	Glycosyl transferases group 1	1.0

Supplementary table S5. Network topological analyses result of hot springs vOTUs monopartite network.

Parameter	Value
Clustering coefficient	0.501
Connected components	233
Network diameter	5
Network radius	1
Network centralization	0.012
Shortest paths	3448
Characteristic path length	1.662
Average number of neighbors	2.418
Number of nodes	784
Network density	0.003
Network heterogeneity	0.717
Isolated nodes	0
Number of self loops	0
Multi-edge node pairs	0

GENERAL DISCUSSION

Studies of viral communities in hot springs have been mostly circumscribed to hypertermophilic or acidic hot springs, and usually focused only in the lytic populations. Therefore, the study of viruses from thermophilic phototrophic microbial mat communities remains largely unexplored except for a few studies providing limited information on viral presence within these communities (Davison et al., 2016; Heidelberg et al., 2009).

Hot springs are discontinuous environments that can be considered as "hot islands" surrounded by a "cold ocean" providing a unique study model in which to evaluate the relevance of environmental factors and dispersal limitation in the establishment and development of viral communities. Microbial communities from these high temperature habitats are mostly dominated by few types of microorganisms (some phyla), and usually are less diverse than lower temperature freshwater habitats, and oceans. In this sense, this simplified natural communities can be an excellent scenario in which to study lytic and lysogenic dynamics of the viral component.

Chile, and in particular the active volcanic region of the Patagonian fjords provide us with a natural laboratory where to study the viral communities of the phototrophic mats that grow in the many hot springs hidden among the deep vegetation of this unique area on the planet.

So far, this is the first time that a study has characterized the composition and activity of the principal viruses of these communities (Guajardo-Leiva et al., 2018), identifying the first

complete genome of a virus infecting thermophilic cyanobacteria (*Fischerella*) in the model hot spring of Porcelana. Although this first study (Chapter 1 of this thesis) gives us a first look at the lytic thermophilic viral communities in phototrophic mats, it was necessary to analyze in depth the rest of the viral community, including its potential hosts and lifestyles. Therefore sequencing the viral enriched fraction was a fundamental task.

Equally important was the comparison between Patagonia viral communities and those communities from other terrestrial hot springs to understand fundamental questions. For example, are these viruses specific to a hot spring or are widely distributed? or, are these viruses affected by environmental factors?.

In the development of this doctoral thesis, many of these questions were answered. However, many of the results obtained opened a larger number of new questions that will need to be answered in the future. So, in these studies we have just laid the first stone in this complex but interesting research field.

1. Caudovirales are active and ubiquitous in the cellular fraction of Porcelana phototrophic microbial mats.

As a first insight into the viral communities in hot springs in Chapter 1, we recovered viral sequences from three cellular metagenomes and metatranscriptomes of Porcelana phototrophic mats at three temperatures (48-58-66 °C). Porcelana microbial mats are mainly built by filamentous representatives of two phototrophic phyla, Cyanobacteria (oxygenic) and Chloroflexi (anoxygenic), with *Fischerella*, *Chloroflexus* and *Roseiflexus* as the main genera, respectively. This data (Guajardo-Leiva et al., 2018) confirmed previous surveys carried out

by our own laboratory (Alcamán-Arias et al., 2018; Mackenzie et al., 2013). Similar results were also recorded in Yellowstone National Park studies in White Creek, Mushroom and Octopus hot springs, that present similar pH, thermal gradient and low sulfide concentrations (Bolhuis et al., 2014; Goldsmith and Miller, 2009; Inskeep et al., 2013; Klatt et al., 2013).

In Chapter 1, we showed that Porcelana dominant and active viruses (~70% and ~68% of metagenomic and metatranscriptomic reads, respectively) belong to the families Myoviridae, Podoviridae and Siphoviridae within the Caudovirales order (Chapter 1, Figure 2), which typically infect Bacteria and some non-hyperthermophilic Archaea (Ackermann and Maniloff, 1998). These results were also supported by TEM images (Chapter 1, Figure 1).

Dominance by Caudovirales was also reported recently from the Brandvlei hot spring, South Africa, a slightly acidic (pH 5.7) hot spring with a moderate temperature (60 ºC) and green microbial mat patches (Zablocki et al., 2017). Previously, in moderate thermophilic phototrophic mats of Yellowstone, the presence of this viral order was suggested only by using indirect genomic approaches as spacers at CRISPR loci of dominant bacterial members (Davison et al., 2016; Heidelberg et al., 2009) or classifications based on nucleotide motives from viral metagenomic data (Davison et al., 2016; Pride and Schoenfeld, 2008).

All this evidence suggests that the Caudovirales order would be abundant in hot springs with similar physicochemical conditions.

Virus-host inference in Porcelana phototrophic mats (Chapter 1, Figure 3B), demonstrated that the most frequent targets for viral infections of Caudovirales were also the most transcriptionally active and abundant components of the bacterial communities, following the KtW model (Breitbart and Rohwer, 2005). In Porcelana at 48 °C and 58 °C, cyanophages were

among the most active viruses (Chapter 1, Figure 3B), as were their cyanobacterial hosts, such as *Fischerella* spp. The latter, are relevant and active members of the community in these mats as exemplified in terms of primary production and nitrogen fixation (Alcamán-Arias et al., 2018; Alcamán et al., 2015). The presence of cyanophages has been previously suggested in Yellowstone hot springs phototrophic mats (Davison et al., 2016; Heidelberg et al., 2009), and more recently in the Brandvlei hot spring, South Africa (Zablocki et al., 2017). In Brandvlei a 10 kb partial genome of a new cyanophage (BHS3) was reconstructed from a viral matagenome, stating that cyanophages appear to be the dominant viruses in this hot spring. The BHS3 contig contains 9 ORFs, with the majority of the identified proteins having a close relation to the Cyanophage PP and *Phormidium* phage Pf-WMP3, which infect freshwater filamentous cyanobacteria *Phormidium* and *Plectonema* (Zablocki et al., 2017).

Caudoviruses were also prevalent at 66°C in Porcelana, in this case potentially associated with Firmicutes, Proteobacteria, and Actinobacteria. These bacterial phyla have also been previously identified in other hot springs at temperatures above 76 °C, such as in Octopus and Bear Paw (Pride and Schoenfeld, 2008).

Even when the phylum Chloroflexi was dominant in the phototrophic mat at 66°C (Chapter 1, Figure 3A), viral sequences related to this taxon could not be retrieved, as neither viruses or viral sequences have been confirmed to infect members of this phylum in any environment. In fact, Davison et al, (2016), described viral contigs associated with *Roseiflexus* sp. from a viral metagenome from Octopus Spring, but unfortunately, only raw reads are publicly available, without taxonomic assignation (Davison et al., 2016).

2. Viral mining of Porcelana metagenomes reveals a new lytic thermophilic cyanopodovirus lineage associated with *Fischerella* spp..

Because of the high abundance of viral sequences associated with cyanophages in the unassembled cellular metagenomes of Porcelana, we assembled each metagenome to interrogate them about the presence of full viral genomes. This analysis recovered a unique complete genome (TC-CHP58) represented by a contig with a typical size of the Podoviridae family (~50 kb). The genome size and viral core proteins close to the Podovirus family point TC-CHP58 as the first full genome of a thermophilic cyanopodovirus. Moreover, the genome organization (Chapter 1, Figure 4B) shows a consistent synteny with other cyanopodoviruses inside the T7 supergroup such as Pf-WMP4, Pf-WMP3, Cyanophage PP, *Anabaena* phage A-4L (Liu et al., 2008; Ou et al., 2015; Zhou et al., 2013), as well as to the recently reported partial genome of the thermophilic BHS3 cyanophage (Zablocki et al., 2017).

The phylogenetic position of TC-CHP58, based on DNA polymerase I (DNApol) (Chapter 1, Figure 6) and Major capsid (MCP) (Chapter 1, Supplementary Figure S4) predicted proteins, confirmed the affiliation of this new virus within the family Podoviridae. This analysis also suggested that TC-CHP58 probably form part of a new genus with other freshwater cyanopodoviruses (Pf-WMP4, Pf-WMP3, Cyanophage PP, and phage A-4L). That new genus will include BHS3 and TC-CHP58 as representatives of a novel, and potentially globally distributed thermophilic cyanophage lineage.

For the new cyanopodovirus TC-CHP58, was possible to verify that *Fischerella* spp. was its putative host, via the analysis of CRISPR spacers found in the cyanobacteria recovered from contigs obtained in the same metagenomic datasets (Chapter 1).

Observations from the CRISPR loci overall temperatures (Chapter 1, Table 2) in Porcelana indicated that, in general, proto-spacers in the TC-CHP58 genome were distributed on coding, and therefore more conserved regions. It has been demonstrated that viruses rapidly escape from CRISPR through mutation (Shmakov et al., 2017). Consequently, there is a greater pressure for negative selection in sectors that code fundamental functions for the success of the viral infection. The expression of 7 different CRISPR loci at different temperatures (Chapter 1, Figure 5), demonstrated the activity of the *Fischerella* spp. defense system against TC-CHP58 over the temperature gradient. Despite variations in the number of CRISPR loci observed at each temperature, with 60% of the total CRISPR loci found in *Fischerella* contigs at 58 ºC, the abundance of CRISPR reads agreed with the abundance of other genes required by these cyanobacteria, such as the RUBISCO gene (Chapter 1, Figure 5). This further verified that the CRISPR loci are from *Fischerella* populations. The different CRISPR loci found over the different temperatures in Porcelana (Table 2), reinforces the notion that diversification of *Fischerella* is partly due to selective pressure exerted by the predation of viruses, such as TC-CHP58. This theory has been previously put forward for *Fischerella thermalis* in Yellowstone (Sano et al., 2018), and also proposed for marine cyanobacteria (Kashtan et al., 2014; Rodriguez-Valera et al., 2009).

The selection pressure of multiple spacers in *Fischerella* CRISPR loci leads to the emergence of single nucleotide variants (SNVs) in the TC-CHP58 viral populations (Chapter1, Table 2), which cause mismatches between spacers and proto-spacers, resulting in the attenuation or evasion of the host immune response (Shmakov et al., 2017). Microorganisms can still use mismatched spacers for interference and/or primed adaptation however, the degree of

tolerance to mismatches for interference among the CRISPR-Cas, varies substantially between different CRISPR-Cas type systems (Shmakov et al., 2017). The variable frequency (0.6 to 0.02) of the corresponding spacer SNVs alleles on TC-CHP58 proto-spacers, suggests that some variants are more prevalent throughout the viral population, regardless of whether the SNV causes a silent mutation. Based on this evidence, it has been proposed for other microbial communities that, only the most recently acquired spacer can exactly match the virus. This suggests that community stability is driven by compensatory shifts in host resistance levels and virus population structure (Andersson, 2008).

3. Lysogens in Porcelana phototrophic microbial mats.

The results shown in the first chapter evidenced that, most abundant viruses in the phototrophic mats of Porcelana were part of the Caudovirales order and specifically that cyanophages were the most active viruses in this system. However, with the analyses performed in Chapter 1, we were only able to capture sequences from isolated viruses associated with sequences in databases. Therefore in the second chapter, we analyzed the viral community in depth, including lytic and lysogenic viral groups, using genome-resolved metagenomics (MAGs) and viral metagenomics approaches to analyze lysogens and the viral communities.

Temperate bacterial viruses (prophages) can enter in symbiosis with their host organism, forming a new unit called lysogen (Knowles et al., 2017). It has been argued that lysogens have a high frequency (20 to 60%) within bacterial cultured strains and represent 66% of the viral genomes in reference databases (McNair et al., 2012; Miller and Day, 2008).

Lysogeny has been historically estimated by quantifying the viral progeny after prophage induction by the mitomycin C (MitC) treatment, a DNA damaging agent (Howard-Varona et al., 2017; Kim and Bae, 2018; Knowles et al., 2017). However, this methodology of induction has been mostly used in cultivable bacteria and its effects in natural communities are not well understood (Kim and Bae, 2018; Knowles et al., 2017). Hence, the identification and quantification of lysogens in mixed natural communities have been even more challenging, than in pure cultures.

In order to solve culture limitations and study the lysogens present in the natural phototrophic mat community of Porcelana, a genome-resolved metagenomic analysis (MAGs) was used. We used 34 MAGs recovered from three previously published cellular metagenomes of Porcelana (Alcamán-Arias et al., 2018; Alcorta et al., 2018; Guajardo-Leiva et al., 2018) that meet the completeness and contamination standards (≥ 90% genome completeness and ≤ 10% contamination) developed by the Genomic Standards Consortium (Bowers et al., 2017). Besides the high quality of the bacterial genomes recovered, the presence of active lysogens was low. We discovered one unique temperate virus in MAG Nº30 (Burkholderia GJ-E10), whereas most of the viral sequences found, corresponded to non-active vestigial prophages (Chapter 2, Table 1). These vestigial sequences lack known lysogeny markers such as integration enzymes, recombination enzymes, tRNAs, and attachment sites (Canchaya et al., 2003). However, these defective prophage sequences have been reported to provide adaptive functions to bacteria such as, gene transfer agents (GTAs) that stochastically transfer fragments of chromosomal DNA to another microorganism, also type 6 secretion systems

(T6SSs) and bacteriocins that are involved in bacterial defense and competition mechanisms (Bobay et al., 2014).

4. Database dependent analyses of composition, lysogenic markers, and prophages in Porcelana viral metagenomics.

In hot springs, lysogeny has been proposed as an effective lifestyle (Breitbart et al., 2004; Pride and Schoenfeld, 2008; Sharma et al., 2018). Breitbart *et al*, in 2004, showed that an experimental mitomycin C induction in California hot springs (74-82 °C) produced an increase of 1.2 to 1.4 fold in the number of VLPs after mitomycin C induction. Schoenfeld *et al*, in 2008 recovered 86 integrases genes in two viral metagenomes obtained from YNP hot springs (74-93 °C). And finally, Sharma *et al*, in 2018 recovered 66 viral genomes from cellular metagenomes of a Himalayan hot spring (50-98 °C), where 47% of the recovered viruses were found to be lysogenic.

In the second chapter of this thesis, in order to better understand natural and mitomycin C induced viral communities we applied viral metagenomics to *in-situ* induction experiments performed in Porcelana phototrophic mats.

Analyses of the predicted proteins in viral metagenomes separated the natural from the MitC induced viral communities (Chapter 2, Figure 1B), however, both communities were dominated by Caudovirales order. Natural viral communities were rich in viruses from Podoviridae family (60-71%), as previously reported in the same hot spring (60-71%), only previously reported as an important component in the same hot spring (Guajardo-Leiva et al., 2018) as well as in Brandvlei hot spring (South Africa) (Zablocki et al., 2017, 2018).

Mitomycin induced communities were in turn dominated by Myoviridae family viruses as well as by environmental or unclassified viruses. Myoviridae family includes known lysogenic lifestyle genera such as Mu-like viruses (*Muvirus*) and some P2-like viruses (*P2virus*) (Canchaya et al., 2003) which can explain the high presence of this family in the induced community.

The search for lysogenic markers has been used as a strategy to study and discover lysogens and prophages in natural communities and isolated bacteria (Howard-Varona et al., 2017). Genes such as integrases and ParA/B (Emerson et al., 2012) or integrases and cI-type repressors (McDaniel et al., 2008) have been used in hypersaline and marine ecosystems respectively, to study lysogenic seasonal dynamics. Besides, databases from lysogenic markers and common viral genes have been used more widely to identify prophages in microbial genomes and metagenomic data (Arndt et al., 2016; Reis-Cunha et al., 2017; Zhou et al., 2011).

In Porcelana, lysogenic markers genes (Chapter 2, Supplementary Table S2) and other genes present in lysogenic viruses (Chapter 2, Supplementary Figure S2) were annotated and commonly found in both natural and MitC induced communities. Moreover, some of these gene markers (e.g. integrases, ParB and recombinases) were found to be more abundant in natural than in the MitC induced communities (Chapter 2, Supplementary Figure S2), suggesting once more that lysogeny is a common feature in natural microbial communities of hot springs (Breitbart et al., 2004; Pride and Schoenfeld, 2008; Sharma et al., 2018).

Additionally, the differential abundance analyses between natural and MitC induced communities, allowed us to detect 12 vPCs that showed statistically significant differences in

the MitC induced communities. Functional annotations of this vPCs did not correspond to any lysogenic markers (Chapter 2, Supplementary Figure S2). However, two interesting protein families arise from this analysis, FtsK/SpoIIIE family (PF01580) and Tubulin/FtsZfamily (PF00091). Those two families were previously described in lytic and lysogenic viruses infecting *Bacillus (Grose et al., 2014)* or *Clostridium* and *Pseudomonas (Kraemer et al., 2012)*.

This differential abundance analyses suggest that MitC treatment has a strong and quantitative effect over a specific group of hosts, those who in turn harbor specific prophages that contain new unknown functions (vPCs) that have not been described in hot springs before.

Additionally, in a genomic context, marker-based (PHASTER) and differential abundance analyses of vOTUs revealed 23 putative lysogenic viruses, which also includes the prophage PP_Burkholderia_GJ-E10 (Chapter 2, Supplementary Figure S3). These 23 vOTUs would mostly infect bacteria of the Firmicutes (14) and Proteobacteria (8) phyla, whereas only Firmicutes viruses showed an increment of their relative abundances after mitomycin treatment of microbial mat at both sites investigated (P50 and P55). Interestingly these two phyla harbor the vast majority of prophage sequences in databases (Canchaya et al., 2003) what has also been described in oceans (McDaniel et al., 2008) and hot springs (Sharma et al., 2018). A metagenomic survey of marine phage integrases showed most of them were closely related to viruses of Proteobacteria (Vibrio) and Firmicutes (Clostridium) and were widespread in GOS samples as well as in Tampa Bay MitC induction experiments (McDaniel et al., 2008). Moreover, metagenomic recovery of viral genomes in a Himalayan hot spring

showed that also Proteobacteria and Firmicutes associated phages composed 28 of the 31 lysogenic viruses found in these microbial mats and sediments (Sharma et al., 2018).

In Porcelana viral communities, lysogenic viruses represented a small fraction of the total number of genomes recovered (vOTUs), and most of them were part of the less abundant genomes (< 1 % of abundance) (Chapter 2, Table 6). However, there was one exception (vOTU P55_C_19) that corresponded to the most abundant (3%) virus in Porcelana (Chapter 2, Table 6), with the highest presence in the natural community at P50 site and in the MitC induced community from P55 site. This vOTU shared 10% of their protein sequences with viruses infecting Firmicutes of the *Streptococcus* and *Clostridium* genera.

Together, our results were in accordance with previous studies in hot springs and other environments pointing Proteobacteria and Firmicutes as most common putative hosts of lysogenic viruses in environmental communities. However, temperate phages richness in Porcelana seems to be lower than in the recent report of viruses from Himalayan hot springs (Sharma et al., 2018). This can be explained due to the fact that the dominant bacterial groups in that hot spring are exactly Proteobacteria and Firmicutes phyla (Sharma et al., 2018), and although in Porcelana these groups are also important, they are far from being the dominant ones (Alcamán-Arias et al., 2018; Guajardo-Leiva et al., 2018).

In addition, Firmicutes lysogens were more susceptible to MitC induction than Proteobacteria members and Proteobacteria lysogens together with a minor number of Firmicutes lysogens were spontaneously induced in Porcelana hot springs. It is important to emphasize that only these two groups were associated with lysogenic viruses regardless of the induction factor that leads to the development of a lytic cycle.

5. Lytic and lysogenic viral networks reveal new, and ecologically relevant viral genera in Porcelana communities

Monopartite protein shared networks are able to group together different viral genomes (vOTUs) into Viral Clusters (VCs) which can be considered a genus with an 80% of precision (Bolduc et al., 2017; Paez-espino et al., 2019).

Using this methodology we constructed a network of Porcelana viral communities and RefSeq-ICTV reference viral genomes (Chapter 2, Figure 5) that showed a higher degree of modularity (3.6 fold), when compared to the network formed by only the RefSeq (1964) reference genomes (Bolduc et al., 2017). This larger modularity occurs due to the fact that most of the Porcelana genomes (369) grouped together forming 103 new viral genera. This high rate of new genera discovery here is even higher than that reported for the marine environment (Jang et al., 2019). This is evident when we compare Porcelana results to those of the Global Ocean Sampling campaign (GOS), where the latter had 15280 viral genomes and only 919 represented new genera (Jang et al., 2019).

Most of these new genera discovered in Porcelana will probably remain completely unknown because most of them (94) lack any information related to taxonomy or host. From these new genera, we identified nine infecting *Fischerella* (2 genera), *Chloroflexus* (1 genus), *Roseiflexus* (1genus), *Meiothermus* (4 genera), *Burkholderia* (1 genus) and *Pedosphaerales* (1 genus) (Chapter 2, Figure 6). The first four new viral taxa infect three genera of bacteria that are known to play an important function in Porcelana microbial mats, as primary producers and builders of these communities (Alcamán-Arias et al., 2018; Alcamán et al., 2015). Therefore, it is possible that these specific host-virus relationships are relevant to the entire

microbial community. This is particularly true in the case of *Fischerella*, on which the rest of the community depends, due to its important ecological role as carbon and nitrogen fixer in this and many other hot springs over the world, since two of the new genera that infect this organism are among the most abundant viral genomes recovered from Porcelana (Chapter 2, Table 6). The first *Fischerella* infecting viral genus, with 5% of total abundance, belongs to the Podoviridae family, and have one full representative genome (TC-CHP58) that was recovered from Porcelana cellular metagenomes in Chapter 1 of this thesis (Guajardo-Leiva et al., 2018). The second viral genus (Chapter 2) belongs to the Siphoviridae family and represent ~ 2% of the total viral community. However, the genomes of this new genus do not resemble any other known cyanophage. In that sense, our results emphasize the need to do a more extensive work in this type of environments, in order to recover a major number of complete genomes of these and other viruses infecting key organisms in this type of worldwide distributed hot springs.

6. Insights into viral communities structure in global hot springs.

In the first and second chapter of this thesis, we have explored the viral communities of Porcelana hot spring, demonstrating their activity and how they interact with their hosts in a natural system. We also described new genomes and discovered new viral genera that could have global implications because some of them infect bacterial groups that are ecologically relevant and common in phototrophic mats of hot springs around the world. According to that, we considered of much relevance to compare the local viral community found in Porcelana to the viral community of other hot springs in Patagonia such as Cahuelmó hot spring, as well as

in a more global scale compare to any other hot springs in the world using public viral metagenomes (YNP and South Africa) with different physicochemical properties. In this way, that analysis was expected to answer fundamental questions to better understand the ecology of viral communities in these extreme systems.

It is well known that colonization of hot springs by microbial phototrophic mats not only has temperature limits but also depends on pH. The upper temperature limit estimated for oxygenic phototrophs is near to 74˚C (Brock, 1973), but Cyanobacteria (the main oxygenic phototrophic phyla in many hot springs mats) generally do not live in habitats below pH 4.5-5. Consistently, there is a phototrophic limit at pH values < 5, and temperature near to 56˚C which correspond to the upper temperature limit for other microbes such as Cyanidales, diatoms, or photosynthetic proteobacteria (Inskeep et al., 2013).

Moreover, pH also produces dichotomous divisions in hyperthermophilic (72-98 °C) environments. Sites of pH values > 4 are dominated by bacteria from Aquificales and Thermotogales, with the contribution of archaea from Desulfurococcales and Thermoproteales. However, at pH values < 4, Sulfolobales archaea dominate (Inskeep et al., 2013; Menzel et al., 2015). According to the above, it is expected that viral communities at different pH and temperature, have a structure concordant with the hosts that inhabit in each particular hot spring (Gudbergsdóttir et al., 2016).

In the present study, a database dependent analysis on a global scale of proteins predicted from viral metagenomes separated the different viral communities into those dominated by archaeal viruses and those of dominated by bacterial viruses (Chapter 3, Figure 1B). As in previous studies, acidic-hyperthermophilic hot springs such as Nymph Lake and also Crater Hill Geyser

(Bolduc et al., 2012, 2015), were enriched in viruses sequences from archaea infecting members of the Sulfolobales from Rudiviridae, Lipothrixviridae and Bicaudaviridae families (Bolduc et al., 2012, 2015).

Also similar to previous studies, circumneutral-hyperthermophilic sites, were dominated by both, unclassified environmental bacterial viruses, and archaeal viruses related to Globuloviridae family and the Hypherthermophilic Archaeal Virus 1 (HAV1) which infect Thermoproteales order (Pride and Schoenfeld, 2008). The last group of circumneutral-thermophilic hot springs, were rich in bacterial viruses from Podoviridae family and unclassified bacterial viruses. The high abundance of Podoviridae sequences was especially interesting since only a few and recent studies have reported this family in hot springs (Guajardo-Leiva et al., 2018; Zablocki et al., 2017, 2018).

In this study, BSL was the only representative of acidic-thermophilic hot springs, and consequently, it was not surprising that it has a unique combination of viral populations (Chapter 3, Figure 1B). This sample was rich in archaeal viruses from Rudiviridae, Bicaudairidae, and Turriviridae families, but also in unclassified ssDNA viruses, environmental viruses such as BSL-RDHV and know representatives of Circoviridae family. The Circoviridae related sequences were the only group of viruses previously reported in BSL (Diemer and Stedman, 2012).

Our database dependent analyses of viral communities on a global scale showed that these communities were structured according to the structure of the host communities that in turns are dependent on two main physicochemical conditions pH and temperature. This natural

clustering of the viral communities defined three main groups of samples: YNP Acidic, YNP Neutral, and Patagonia Phototrophic.

7. Role of physicochemical factors in the protein universe of global hot springs.

The use of database dependent analysis in unexplored ecosystems only allow us to obtain a rough idea of the structure of viral communities, since taxonomy and abundances of each taxon are obtained by comparison to a reference and not between the existing universe of sequences from the samples investigated. In order to avoid this bias, we decided to use vPCs as an approach to calculate the diversity (Shannon Index) and evenness (Pielou's) of viral communities (Chapter 3, Supplementary Figure S1). A moderate correlation was found for the alpha diversity and evenness with pH and temperature (Chapter 3, Supplementary Figure S2) which shows that both physicochemical factors have a predictive effect on viral communities in these hot springs. These findings are in agreement with the cellular counterpart diversity reports from hot springs around the world, where a moderate correlation with pH (0.4-0.44) (Power et al., 2018; Sharp et al., 2014) and a stronger correlation with temperature (0.79) (Sharp et al., 2014), was found.

Therefore, and even with the obvious limitations of the reduced number of hot spring viral metagenomes available for analyses, our results strongly suggest that viruses have diversity and evenness patterns that are affected by pH and temperature, as it occurs with the host counterpart.

Hot springs viral protein universe recovered in the present study was hardly partitioned in three clusters (Chapter 3, Figure 3A), which coincide with the clusters defined before by the

divisions generated by the temperature and pH that in turn divide the communities of hosts (YNP Acidic, YNP Neutral, and Patagonia Phototrophic). Ordination of vPCs (Chapter 3, Figure 3B) confirms the results of the hierarchical clustering of viral communities (Chapter 3, Figure 3A) even when only ~ 36% of the variance was explained by PCo1 and PCo2. Stronger predictors ($p<0.05$) of the viral structure were the pH and latitude environmental variables, whereas temperature can be considered weaker (because of arrows lengths, Chapter 3, Figure 3B). Categorical variable, sample source, also have a strong effect on the ordination, particularly on phototrophic microbial mats from Patagonia (P50, P55 and CA) and sediment sample from BSL in California. Permanova analyses (Chapter 3, Table 2) corroborate the explanatory role of pH (12% of the total variance) in the dissimilarity of viral communities structure, even when it is not statistically supported ($P \leq 0.0735$).

Together, these results showed an organization of the viral communities in the physicochemical gradients across the samples, and a different viral protein repertory at each hot spring, suggesting that geographic location (latitude) and pH are the factors that more influenced the viral community structure. In particular, latitudinal biogeographic patterns have been extensively studied in macroorganism (Hillebrand, 2004) and also been described in marine (Fuhrman et al., 2008) and terrestrial microorganisms (Andam et al., 2016; Sharp et al., 2014). Viral counterpart has also shown latitudinal influence on community structure in marine datasets from the Pacific Ocean Virome (POV) (Hurwitz et al., 2014) and Tara Ocean Virome (TOV) (Brum et al., 2015). However, the mechanism underlying these latitudinal patterns are still under debate for viral communities. Here, and due to the nature of vPCs that can measure deep evolutionary events, these patterns are probably the product of historical

events (geological, ecological or demographic), such as glaciations, marine currents, and atmospheric circulation that influenced their dispersion and diversification (Andam et al., 2016).

Regarding pH, it has been previously described as a significant factor influencing the microbial community composition in hot springs (Inskeep et al., 2013; Menzel et al., 2015). It has even been proposed as the main driver of these communities to date (Power et al., 2018). It is also known that pH generates specific adaptations in microorganisms to deal with altered nutrient availability, metal solubility and organic carbon characteristics that result in a reduced number of taxa that can physiologically tolerate these conditions (Power et al., 2018). In this way, pH will affect the composition of the host and consequently the viral communities that prey and reproduce in those environments (niche specialization).

8. Endemism of global hot springs vOTUs is dependent on geographic location and temperature.

Because of vPCs capture deep evolutionary events as well as a community structure more focused on function, we decided to use here vOTUs as a complementary approach. The vOTUs capture evolutionary and ecologically cohesive populations of closely related viral genomes (genotypes), which have no fitness differences in the same niche space (host) (Duhaime et al., 2017). Therefore, vOTUs provide a metric unit to analyze viral communities at genome and population levels.

Alpha diversity and evenness showed a strong and moderate correlation of vOTUs with temperature (Chapter 3, Supplementary Figure S6), concluding that only this factor and not

pH have a predictive effect on viral diversity. As discussed before, a strong correlation of diversity with temperature (0.79) has been found for bacterial and archaeal communities in hot springs (Sharp et al., 2014). The latter supports our findings, despite the small number of the existing samples analyzed here, and strongly suggest that viral diversity and evenness measured at genomes level are affected mainly by temperature.

Hierarchical clustering analyses of vOTUs (Chapter 3, Figure 5A) analyses, congruently separated viral communities in two (YNP Acidic and Patagonia Phototrophic) of the three clusters defined on our previous analyses. The absence of YNP Neutral cluster was related to the impossibility to detect any vOTU in Bear paw and Octopus samples, from the work of Schoenfeld *et al.*, in 2008, mostly due to the lack of coverage (75 % coverage threshold) even when vOTUs assembled from Octopus (Octopus and Octopus G7162 co-assembly) were available.

Distance between viral communities inside each cluster was higher than in the vPCs analyses, because of the strictness of vOTUs detection thresholds and the fast evolving nature of DNA compared to proteins. However, the more permissive nature of vPCs based analyzes is necessary for comparative viral genomics at deeper evolutionary times and therefore not relevant to populations and speciation process (Jun et al., 2009).

In viral populations, speciation is proposed to follow a parapatric model, which implies the existence of incomplete genetic barriers and substantial opportunities to gene flow (Duffy et al., 2007). Notwithstanding, because of viruses high mutation rate or reintroduction of alleles that induce slow adaptation of emerging new viruses lineage in novel hosts, ranges of the closely related ancestral and evolved viruses stop to overlap (Duffy et al., 2007). The last

scenario could lead to the genetic discontinuity between nearby hot springs and even inside the same hot springs over the thermal gradients. This feature was observed in our analyses, which implies that even when gene flow is not totally interrupted, a local adaptation of host and therefore their viruses could produce the actual structure observed.

Ordination of vOTUs (Chapter 3, Figure 5B) confirms the results of the hierarchical clustering (Chapter3, Figure 5A) showing the same groups of viral communities, even when only half of the variance was explained by PCo1 and PCo2. Environmental variables, stronger predictors ($p<0.05$) of the viral structure, were latitude and temperature, where conversely pH was considered weaker (because of arrows lengths). Categorical variable, defined as sample source, also have a predictive effect on the ordination but was diffuse between samples. However, Permanova analyses (Chapter 3, Table 3) produced no significant results for the main effect of the four predictor factors that we found in our exploratory PCoA analyses, which is not contradictory because the latter explain 100% of the variance observed.

Viral communities studied here showed a high number of exclusive vOTUs (endemism) per viral cluster or even sample (Chapter 3, Supplementary Figure S4) with only one vOTU found in two samples that were not part of the same cluster (Brandvlei and Porcelana at 50 ºC). Endemisms of specific viral groups (Bautista et al., 2017) and also viral communities (Bolduc et al., 2015) has been previously suggested to occur in several YNP hot springs (Bautista et al., 2017; Bolduc et al., 2015). Bautista and colleagues (2017) showed that Sulfolobus rudiviruses have a biogeographic distribution inside the YNP, while Bolduc and colleagues (2015) showed that ~62 % of the viral groups from Crater Hill Geyser and Nymph Lake are common in both samples (Bolduc et al., 2015).

This high endemism contrast with the low endemic vOTUs (15%) pattern found at global scale in the Tara Ocean marine environments dataset (Brum et al., 2015), but it is consistent with more discontinuous environment studies such as those comparing cyanophages populations from the marine coast and offshore (Gregory et al., 2016).

Together, these vOTUs results found here showed that hot springs viral communities were structured across different samples by endemic viruses at different levels (hot spring and clusters) but also by different abundances of common vOTUs. Our analyses suggest that geographic location (latitude) and temperature influences the viral populations (vOTUs) and consequently the community structure.

9. Unprecedented modularity in global hot springs viral network.

As mentioned before, monopartite protein shared networks are able to group together different viral genomes (vOTUs) into Viral Clusters (VCs) which can be considered a genus with 80% of precision. Here, the hot spring viral network showed extremely high modularity with 233 connected components (genera) related to 784 genomes, compared to the previous network study that analyzed 1964 RefSeq genomes of archaeal and bacterial viruses resulting in only 46 modules (Bolduc et al., 2017). In the same study, the most connected component from RefSeq network includes 1891 genomes of the order Caudovirales (Bolduc et al., 2017), while the two most connected components in our hot spring network include only 20 unclassified genomes. However, a close analysis of only the 73 archaeal viruses on Bolduc et al (2007) work, show that they are divided into 8 connected components, which implies that 3.7% of all viruses occupy 17.4% of all connected components (Bolduc et al., 2017). The latter reveals a common pattern of a high degree of modularity in viral communities from extreme ecosystems

since most of the genomes from archaeal viruses come from thermophilic or hypersaline systems.

This high modularity found suggests a little or no genetic exchange between viral genomes of different genera, or in empirical words that they do not share enough proteins to establish a statistically significant relationship between their genomes. Two plausible explanations arise here for this phenomenon. First, protein based analyses reveal deep and ancient evolutionary processes, so the different lineages that we see today diverged a long time ago from a common ancestor (Gregory et al., 2016). Second, in hot springs exist a tight co-evolution between viruses and their hosts, that acts as a barrier to gene flow, which in viruses is strictly restricted to episodes when two viruses co-infect the same host, requiring spatial proximity and also shared a host range (Gregory et al., 2016). The last scenario was proposed here for cyanophages populations that infect the filamentous cyanobacteria *Fischerella* in Porcelana hot spring (Patagonia, Chile), were susceptible and resistant host populations across a 40 meters thermal gradient select specific cyanophages populations through CRISPR-Cas immunity (Chapter 1) (Guajardo-Leiva et al., 2018).

Moreover, host assignment to viral genera (Chapter 3, Figure 6) showed that some specific hosts have more than one viral genus associated and also that each genera is composed by a large number of genomes (vOTUs), which in fact represent different viral populations. The viral richness found, associated with each host, shows the existence of specific host and virus pairs capable of maintaining a high local richness. Especially interesting are the viral populations that infect common members of phototrophic microbial mats such as *Fischerella*,

Cloroflexus and *Meiothermus*, which have very low representation even in the large environmental databases such as IMG/VR.

Another interesting information obtained from the network analysis is that each genus is composed mainly of genomes of a single thermal system, genomes of nearby thermal systems (viral community clusters) or only by IMG/VR genomes. Exceptionally, there are some genera that combined genomes of Brandvlei (South Africa) with genomes of Porcelana and Cahuelmó (Chile), and finally genus with IMG/VR genomes mixed with those of environmental samples. This demonstrates the existence of a biogeographic effect in the distribution of the different genera, with a few cosmopolitan genera but most of them unique for each locality.

The relationship between Porcelana and Brandvlei viral genomes was also evidenced by the fact that this samples shared some vPCs and at least one vOTU, which shows a potential genetic flow between these two hot springs separated by 7672 km and the Atlantic Ocean between them. This demonstrated once more, that there is a transport of viral particles, possibly by winds from Africa to South America (Griffin, 2007) and that this transport has been maintained over time. The latter, together with the existence of Porcelana hosts susceptible to migrant and local viruses could have maintained the cohesion between the viral populations of these two hot springs.

In the present thesis, it has been demonstrated that viral communities of thermophilic phototrophic mats are genetically diverse, and structured in a classic log normal distribution, dominated by approximately 21 genotypes out of more than 800 detected. These viral communities are diverse and unique, as compared against the most complete databases

available, i.e. IMG/VR, and other hot springs viral metagenomes across the globe. The latter evidence both the presence of untapped genetic diversity and the under sampling of these microbial consortia. We also observe a long tail of rare viral taxa that were present in half of the microbial mat samples, which might suggest that they have relevant functional roles in these microbial communities.

Finally, our results strongly suggest that the dominant viruses of these microbial communities have an impact on the diversity and evolution of the most active and abundant cellular hosts such as the cyanobacteria from genus *Fischerella*. In this primary producer, the presence of a differential number and diversity of CRISPR spacer sequences in different *Fischerella* MAGs evidenced a viral-driven host genotype richness and diversification (i.e. the number of hosts with different CRISPR arrays). Host diversity is a key determinant of pathogen spread where population-level CRISPR spacer diversity can limit phage persistence. However, in these natural thermal systems, the existence of multiple different genera of phages associated with the same host leads to immune system evasion not only by mutation but also by viral recombination. The last, confirm the validity of the Red Queen hypothesis in hot springs, that posits that antagonistic coevolution between interacting species results in recurrent natural selection via constant cycles of adaptation and counter-adaptation.

GENERAL CONCLUSIONS

Viral communities of phototrophic microbial mats in hot springs differ from those that inhabit thermal waters of high temperature > 73°C, acid pH or combination of both parameters. The viral component of these phototrophic communities is dominated by viruses of the Caudovirales order that mainly infect Bacteria. In contrast, higher temperature and acidic pH environments are usually dominated by viruses from different families and genera that infect mainly Archaea.

Porcelana hot spring in particular was dominated mostly by new and unclassified viral genus and cyanophages of Podoviridae and Siphoviridae families. These cyanophages infected *Fischerella* the most active and abundant genus of cyanobacteria in the microbial mats. Consequently cyanophages were the most active group of viruses in Porcelana phototrophic mat. The high abundance of cyanophage sequences in the viral fraction as well as the high number of transcripts associated to this group allow us to infer that this heterogeneous group of viruses have a lytic lifestyle. In the same line, unclassified viruses that infect other active and abundant hosts in Porcelana microbial mats such as *Chloroflexus*, *Roseiflexus* and *Meiothermus* also presented a lytic lifestyle. In contrast, viral groups that infect heterotrophic bacteria of the phyla Firmicutes and Proteobacteria presented a lysogenic lifestyle.

The transition to a lytic infectious cycle in the viruses associated with Firmicutes could be induced by mitomycin C, however the lysogenic viruses associated with Proteobacteria were

naturally induced in Porcelana. Consequently, we propose that the ecological model “kill the winner” is the one that best fits the viral communities of the active primary producers, while “piggyback the winner” adjust better to viral communities of the heterotrophic bacteria that take advantage of the primary production in thermophilic phototrophic microbial mats, such as those of Porcelana.

Additionally the assembly and description of the first full genome of a thermophilic cyanophage (TC-CHP58) allowed us to explore the mechanisms of interaction between this new cyanophage and its natural host *Fischerella*. This was reflected in the active coevolution of the thermophilic cyanophage TC-CHP58 and its host *Fischerella* where the first suffered a diversification of its population through single nucleotide variants (SNV), and *Fischerella*, through CRISPR heterogeneity following the Red Queen hypothesis.

Finally the discontinuous nature of hot springs environments provide us a unique model in which to evaluate the relevance of environmental factors and dispersal limitation in the establishment and development of viral communities. Our analyses revealed a biogeographic pattern, suggesting that viruses can be passively transported by air, on local scale but probably also on a global scale and then locally structured by environmental conditions (pH and temperature) that affect the host community structure.

As concluding remark, we considered that hot springs are natural laboratories that offer unique characteristics to observe interesting evolutionary and coevolutionary processes in different temporal and spatial scales, as well as untapped sources of microbial biodiversity that treasure a number of proteins and enzymes whose future importance may not be suspected.

REFERENCES

Ackermann, H.-W., and Maniloff, J. (1998). Taxonomy of bacterial viruses: Establishment of tailed virus genera and the order Caudovirales. Arch. Virol. 143, 2051–2063. doi:10.1007/s007050050442.

Alcamán-Arias, M. E., Pedrós-Alió, C., Tamames, J., Fernández, C., Pérez-Pantoja, D., Vásquez, M., et al. (2018). Diurnal changes in active carbon and nitrogen pathways along the temperature gradient in porcelana hot spring microbial mat. Front. Microbiol. 9, 1–17. doi:10.3389/fmicb.2018.02353.

Alcamán, M. E., Fernandez, C., Delgado, A., Bergman, B., and Díez, B. (2015). The cyanobacterium Mastigocladus fulfills the nitrogen demand of a terrestrial hot spring microbial mat. ISME J. 9, 2290–2303. doi:10.1038/ismej.2015.63.

Alcorta, J., Espinoza, S., Viver, T., Alcamán-Arias, M. E., Trefault, N., Rosselló-Móra, R., et al. (2018). Temperature modulates Fischerella thermalis ecotypes in Porcelana Hot Spring. Syst. Appl. Microbiol. 41, 531–543. doi:10.1016/j.syapm.2018.05.006.

Andam, C. P., Doroghazi, J. R., Campbell, A. N., Kelly, P. J., Choudoir, M. J., and Buckley, D. H. (2016). A Latitudinal Diversity Gradient in Terrestrial Bacteria of the Genus Streptomyces. MBio 7, 1–9. doi:10.1128/mBio.02200-15.

Andersson, A. F. (2008). Virus Population Dynamics and Natural Microbial Communities. Science (80-.). 1047. doi:10.1126/science.1157358.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., et al. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res. 44, W16–W21. doi:10.1093/nar/gkw387.

Arroyo M. T. K., Marquet P. A., Marticorena C., Simonetti J. A., Cavieres L., Squeo F., Rozzi R. 2004. Chilean winter rainfall-Valdivian forests. Hotspots Revisited: Earth's Biologically Wealthiest and Most Threatened Ecosystems. 99-103.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., and Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc. Natl. Acad. Sci. 108, 17486–17491. doi:10.1073/pnas.1110889108.

Bautista, M. A., Black, J. A., Youngblut, N. D., and Whitaker, R. J. (2017). Differentiation and structure in Sulfolobus islandicus rod-shaped virus populations. Viruses 9, 1–15. doi:10.3390/v9050120.

Bell, E. M. (2012). “Alkaline environments.,” in Life at extremes: environments, organisms and strategies for survival, ed. E. M. Bell (Wallingford: CABI), 380–401. doi:10.1079/9781845938147.0380.

Bergh, Ø., BØrsheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. Nature 340, 467–468. doi:10.1038/340467a0.

Bhaya, D., Grossman, A. R., Steunou, A. S., Khuri, N., Cohan, F. M., Hamamura, N., et al. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. ISME J. 1, 703–713. doi:10.1038/ismej.2007.46.

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

Bolduc, B., Jang, H. Bin, Doulcier, G., You, Z.-Q., Roux, S., and Sullivan, M. B. (2017). vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. PeerJ 5, e3243. doi:10.7717/peerj.3243.

Bolduc, B., Shaughnessy, D. P., Wolf, Y. I., Koonin, E. V., Roberto, F. F., and Young, M. (2012). Identification of Novel Positive-Strand RNA Viruses by Metagenomic Analysis of Archaea-Dominated Yellowstone Hot Springs. J. Virol. 86, 5562–5573. doi:10.1128/JVI.07196-11.

Bolduc, B., Wirth, J. F., Mazurie, A., and Young, M. J. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. ISME J. 9, 2162–2177. doi:10.1038/ismej.2015.28.

Bolhuis, H., Cretoiu, M. S., and Stal, L. J. (2014). Molecular ecology of microbial mats. FEMS Microbiol. Ecol. 90, 335–350. doi:10.1111/1574-6941.12408.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725–731. doi:10.1038/nbt.3893.

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13, 278–284. doi:10.1016/j.tim.2005.04.003.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. 99, 14250–14255. doi:10.1073/pnas.202488399.

Breitbart, M., Wegley, L., Leeds, S., Rohwer, F., and Schoenfeld, T. (2004). Phage Community Dynamics in Hot Springs These include : Phage Community Dynamics in Hot Springs. Appl. Environ. Microbiol. 70, 1633–1640. doi:10.1128/AEM.70.3.1633.

Brock, T. D. (1967). Life at High Temperatures: Evolutionary, ecological, and biochemical significance of organisms living in hot springs is discussed. Science (80-.). 158, 1012–1019. doi:10.1126/science.158.3804.1012.

Brock, T. D. (1973). Lower pH Limit for the Existence of Blue-Green Algae: Evolutionary and Ecological Implications. Science (80-.). 179, 480–483. doi:10.1126/science.179.4072.480.

Brum, J. R., Sullivan, M. B., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., et al. (2015). Patterns and ecological drivers of ocean viral communities. Science (80-.). 348, 1261498-1–11. doi:10.1126/science.1261498.

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brüssow, H. (2003). Prophage genomics. Microbiol. Mol. Biol. Rev. 67, 238–76, table of contents. doi:10.1128/MMBR.67.2.238.

Cole, J. K., Peacock, J. P., Dodsworth, J. A., Williams, A. J., Thompson, D. B., Dong, H., et al. (2013). Sediment microbial communities in Great Boiling Spring are controlled by temperature and distinct from water communities. ISME J. 7, 718–729. doi:10.1038/ismej.2012.157.

Coman, C., Drugă, B., Hegedus, A., Sicora, C., and Dragoş, N. (2013). Archaeal and bacterial diversity in two hot spring microbial mats from a geothermal region in Romania. Extremophiles 17, 523–534. doi:10.1007/s00792-013-0537-5.

Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C., et al. (2008). Major viral impact on the functioning of benthic deep-sea ecosystems. Nature 454, 1084–1087. doi:10.1038/nature07268.

Davison, M., Treangen, T. J., Koren, S., Pop, M., and Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. PLoS One 11, 1–23. doi:10.1371/journal.pone.0160574.

Diemer, G. S., and Stedman, K. M. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7, 13. doi:10.1186/1745-6150-7-13.

Duffy, S., Burch, C. L., and Turner, P. E. (2007). Evolution of host specificity drives reproductive isolation among RNA viruses. Evolution (N. Y). 61, 2614–2622. doi:10.1111/j.1558-5646.2007.00226.x.

Duhaime, M. B., Solonenko, N., Roux, S., Verberkmoes, N. C., Wichels, A., and Sullivan, M. B. (2017). Comparative omics and trait analyses of marine Pseudoalteromonas phages advance the phage OTU concept. Front. Microbiol. 8, 1–16. doi:10.3389/fmicb.2017.01241.

Duhart P., Crignola P., Ordoñez A., Muñoz J. 2000. Franjas metalogénicas en Chiloé Continental (41°- 44° S), pp 1–5. In: IX Congreso Geológico Chileno.

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B., and Banfielda, J. F. (2012). Dynamic viral populations in hypersaline systems as revealed by metagenomic assembly. Appl. Environ. Microbiol. 78, 6309–6320. doi:10.1128/AEM.01212-12.

Feiner, R., Argov, T., Rabinovich, L., Sigal, N., Borovok, I., and Herskovits, A. A. (2015). A new perspective on lysogeny: Prophages as active regulatory switches of bacteria. Nat. Rev. Microbiol. 13, 641–650. doi:10.1038/nrmicro3527.

Fortey, R.; Pankhurst, R.; Herve, F. 1992, Devonian Trilobites at Buill, Chile (42°S). Revista Geológica de Chile 19 (2): 133-143.

Fuhrman, J. A., Brown, M. V., Green, J. L., Schwalbach, M. S., Brown, J. H., Steele, J. A., et al. (2008). A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. 105, 7774–7778. doi:10.1073/pnas.0803070105.

Gemerden, H. Van (1993). Microbial mats: a joint venture. Mar. Geol. 113, 3–25.
Gobler, C. J., Hutchins, D. A., Fisher, N. S., Cosper, E. M., and Saňudo-Wilhelmy, S. A. (1997). Release and bioavailability of C, N, P Se, and Fe following viral lysis of a marine chrysophyte. Limnol. Oceanogr. 42, 1492–1504. doi:10.4319/lo.1997.42.7.1492.

Goldsmith, C. S., and Miller, S. E. (2009). Modern uses of electron microscopy for detection of viruses. Clin. Microbiol. Rev. 22, 552–563. doi:10.1128/CMR.00027-09.

Gregory, A. C., Solonenko, S. A., Ignacio-Espinoza, J. C., LaButti, K., Copeland, A., Sudek, S., et al. (2016). Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics 17, 930. doi:10.1186/s12864-016-3286-x.

Griffin, D. W. (2007). Atmospheric movement of microorganisms in clouds of desert dust and implications for human health. Clin. Microbiol. Rev. 20, 459–477. doi:10.1128/CMR.00039-06.

Grose, J. H., Jensen, G. L., Burnett, S. H., and Breakwell, D. P. (2014). Genomic comparison of 93 Bacillus phages reveals 12 clusters, 14 singletons and remarkable diversity. BMC Genomics 15. doi:10.1186/1471-2164-15-1184.

Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F., and Díez, B. (2018). Active crossfire between cyanobacteria and cyanophages in phototrophic mat communities within hot springs. Front. Microbiol. 9. doi:10.3389/fmicb.2018.02039.

Gudbergsdóttir, S. R., Menzel, P., Krogh, A., Young, M., and Peng, X. (2016). Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. Environ. Microbiol. 18, 863–874. doi:10.1111/1462-2920.13079.

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T., and 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. doi:10.1371/journal.pone.0004169.

Hewson, I., Chow, C., and Fuhrman, J. A. (2010). Ecological Role of Viruses in Aquatic Ecosystems. Encycl. Life Sci. doi:10.1002/9780470015902.a0022546.

Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. Am. Nat. 163, 192–211. doi:10.1086/381004.

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

Hurwitz, B. L., Westveld, A. H., Brum, J. R., and Sullivan, M. B. (2014). Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. Proc. Natl. Acad. Sci. 111, 10714–10719. doi:10.1073/pnas.1319778111.

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., and Rusch, D. B. (2013). The YNP metagenome project: Environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. Front. Microbiol. 4, 1–15. doi:10.3389/fmicb.2013.00067.

Inskeep, W. P., Rusch, D. B., Jay, Z. J., Herrgard, M. J., Kozubal, M. A., Richardson, T. H., et al. (2010). Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. PLoS One 5. doi:10.1371/journal.pone.0009773.

Jang, H. Bin, Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. doi:10.1038/s41587-019-0100-8.

Jun, S.-R., Sims, G. E., Wu, G. A., and Kim, S.-H. (2009). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. Proc. Natl. Acad. Sci. 107, 133–138. doi:10.1073/pnas.0913033107.

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science (80-.). 344, 416–420. doi:10.1126/science.1248575.

Kim, M. S., and Bae, J. W. (2018). Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J. 12, 1127–1141. doi:10.1038/s41396-018-0061-9.

Klatt, C. G., Inskeep, W. P., Herrgard, M. J., Jay, Z. J., Rusch, D. B., Tringe, S. G., et al. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. Front. Microbiol. 4, 1–23. doi:10.3389/fmicb.2013.00106.

Klatt, C. G., Wood, J. M., Rusch, D. B., Bateson, M. M., Hamamura, N., Heidelberg, J. F., et al. (2011). Community ecology of hot spring cyanobacterial mats: Predominant populations and their functional potential. ISME J. 5, 1262–1278. doi:10.1038/ismej.2011.73.

Knowles, B., Bailey, B., Boling, L., Breitbart, M., Cobián-Güemes, A., del Campo, J., et al. (2017). Variability and host density independence in inductions-based estimates of environmental lysogeny. Nat. Microbiol. 2, 17064. doi:10.1038/nmicrobiol.2017.64.

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

Kraemer, J. A., Erb, M. L., Waddling, C. A., Montabana, E. A., Zehr, E. A., Wang, H., et al. (2012). A phage tubulin assembles dynamic filaments by an atypical mechanism to center viral DNA within the host cell. Cell 149, 1488–1499. doi:10.1016/j.cell.2012.04.034.

Liu, X., Kong, S., Shi, M., Fu, L., Gao, Y., and An, C. (2008). Genomic analysis of freshwater cyanophage Pf-WMP3 infecting cyanobacterium Phormidium foveolarum: The conserved elements for a phage. Microb. Ecol. 56, 671–680. doi:10.1007/s00248-008-9386-7.

Mackenzie, R., Pedrós-Alió, C., and Díez, B. (2013). Bacterial composition of microbial mats in hot springs in Northern Patagonia: Variations with seasons and temperature. Extremophiles 17, 123–136. doi:10.1007/s00792-012-0499-z.

McDaniel, L., Breitbart, M., Mobberley, J., Long, A., Haynes, M., Rohwer, F., et al. (2008). Metagenomic analysis of lysogeny in Tampa Bay: Implications for prophage gene expression. PLoS One 3. doi:10.1371/journal.pone.0003263.

McNair, K., Bailey, B. A., and Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28, 614–618. doi:10.1093/bioinformatics/bts014.

Menzel, P., Gudbergsdóttir, S. R., Rike, A. G., Lin, L., Zhang, Q., Contursi, P., et al. (2015). Comparative Metagenomics of Eight Geographically Remote Terrestrial Hot Springs. Microb. Ecol. 70, 411–424. doi:10.1007/s00248-015-0576-9.

Miller, S. R., Purugganan, M. D., and Curtis, S. E. (2006). Molecular population genetics and phenotypic diversification of two populations of the thermophilic cyanobacterium Mastigocladus laminosus. Appl. Environ. Microbiol. 72, 2793–2800. doi:10.1128/AEM.72.4.2793-2800.2006.

Miller, S. R., Strong, A. L., Jones, K. L., and 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. Appl. Environ. Microbiol. 75, 4565–4572. doi:10.1128/AEM.02792-08.

Miller, R. V., and Day, M. J. (2008). “Contribution of lysogeny, pseudolysogeny, and starvation to phage ecology,” in Bacteriophage Ecology, ed. S. T. Abedon (Cambridge: Cambridge University Press), 114–144. doi:10.1017/CBO9780511541483.008.

Munson-Mcgee, J. H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R. J., Weitz, J. S., et al. (2018). A virus or more in (nearly) every cell: Ubiquitous networks of virus-host interactions in extreme environments. ISME J. 12, 1706–1714. doi:10.1038/s41396-018-0071-7.

O’Malley, M. A. (2008). “Everything is everywhere: but the environment selects”: ubiquitous distribution and ecological determinism in microbial biogeography. Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci. 39, 314–325. doi:10.1016/j.shpsc.2008.06.005.

Ou, T., Liao, X. Y., Gao, X. C., Xu, X. D., and Zhang, Q. Y. (2015). Unraveling the genome structure of cyanobacterial podovirus A-4L with long direct terminal repeats. Virus Res. 203, 4–9. doi:10.1016/j.virusres.2015.03.012.

Paez-espino, D., Roux, S., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2019). IMG/VR v.2.0 : an integrated data management and analysis system for cultivated and environmental viral genomes. 47, 678–686. doi:10.1093/nar/gky1127.

Papke, R. T., Ramsing, N. B., Bateson, M. M., and Ward, D. M. (2003). Geographical isolation in hot spring cyanobacteria. Environ. Microbiol. 5, 650–659. doi:10.1046/j.1462-2920.2003.00460.x.

Pawlowski, A., Rissanen, I., Bamford, J. K. H., Krupovic, M., and Jalasvuori, M. (2014). Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. Arch. Virol. 159, 1541–1554. doi:10.1007/s00705-013-1970-6.

Power, J. F., Carere, C. R., Lee, C. K., Wakerley, G. L. J., Evans, D. W., Button, M., et al. (2018). Microbial biogeography of 925 geothermal springs in New Zealand. Nat. Commun. 9. doi:10.1038/s41467-018-05020-y.

Prangishvili, D., and Garrett, R. A. (2004). Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses (vol 32, pg 204, 2003). Biochem. Soc. Trans. 32, 1133. doi:10.1042/BST0320204.

Pride, D. T., and Schoenfeld, T. (2008). Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. BMC Genomics 9, 1–16. doi:10.1186/1471-2164-9-420.

Rachel, R., Bettstetter, M., Hedlund, B. P., Häring, M., Kessler, A., Stetter, K. O., et al. (2002). Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch. Virol. 147, 2419–2429. doi:10.1007/s00705-002-0895-2.

Redder, P., Peng, X., Brügger, K., Shah, S. A., Roesch, F., Greve, B., et al. (2009). Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. Environ. Microbiol. 11, 2849–2862. doi:10.1111/j.1462-2920.2009.02009.x.

Reis-Cunha, J. L., Bartholomeu, D. C., Earl, A. M., Birren, B. W., and Cerqueira, G. C. (2017). ProphET, Prophage Estimation Tool: a standalone prophage sequence prediction tool with self-updating reference database. bioRxiv, 176750. doi:10.1101/176750.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pašić, L., Thingstad, T. F., Rohwer, F., et al. (2009). Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. 7, 828–836. doi:10.1038/nrmicro2235.

Rohwer, F., and Thurber, R. V. (2009). Viruses manipulate the marine environment. Nature 459, 207–212. doi:10.1038/nature08060.

Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. (2018). Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. Nat. Ecol. Evol. 2, 510–519. doi:10.1038/s41559-017-0435-9.

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., and Mead, D. (2008). Assembly of viral metagenomes from Yellowstone hot springs. Appl. Environ. Microbiol. 74, 4164–4174. doi:10.1128/AEM.02598-07.

Sharma, A., Schmidt, M., Kiesel, B., Mahato, N. K., Cralle, L., Singh, Y., et al. (2018). Bacterial and Archaeal Viruses of Himalayan Hot Springs at Manikaran Modulate Host Genomes. Front. Microbiol. 9, 1–15. doi:10.3389/fmicb.2018.03095.

Sharp, C. E., Brady, A. L., Sharp, G. H., Grasby, S. E., Stott, M. B., and Dunfield, P. F. (2014). Humboldt's spa: Microbial diversity is controlled by temperature in geothermal environments. ISME J. 8, 1166–1174. doi:10.1038/ismej.2013.237.

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., and Koonin, E. V. (2017). The CRISPR spacer space is dominated by sequences from species-specific mobilomes. MBio 8, 1–18. doi:10.1128/mBio.01397-17.

Spiess, F. N., Macdonald, K. C., Atwater, T., Ballard, R., Carranza, A., Cordoba, D., et al. (1980). East Pacific Rise: Hot Springs and Geophysical Experiments. Science (80-.). 207, 1421–1433. doi:10.1126/science.207.4438.1421.

Stern, C. R. (2004). Active Andean volcanism: its geologic and tectonic setting. Rev. geológica Chile 31. doi:10.4067/S0716-02082004000200001.

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

Thiel, V., Wood, J. M., Olsen, M. T., Tank, M., Klatt, C. G., Ward, D. M., et al. (2016). The dark side of the mushroom spring microbial mat: Life in the shadow of chlorophototrophs. I. Microbial diversity based on 16S rRNA gene amplicons and metagenomic sequencing. Front. Microbiol. 7, 1–25. doi:10.3389/fmicb.2016.00919.

Thingstad, T. F., Bratbak, G., and Heldal, M. (2008). "Aquatic phage ecology," in Bacteriophage Ecology, ed. S. T. Abedon (Cambridge: Cambridge University Press), 251–280. doi:10.1017/CBO9780511541483.013.

Torrella, F., and Morita, R. Y. (1979). Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications. Appl. Environ. Microbiol. 37, 774–8.

Wang, H., Yu, Y., Liu, T., Pan, Y., Yan, S., and Wang, Y. (2015). Diversity of putative archaeal RNA viruses in metagenomic datasets of a yellowstone acidic hot spring. Springerplus 4, 1–6. doi:10.1186/s40064-015-0973-z.

Wilhelm, S., and Suttle, C. (2000). Viruses as regulators of nutrient cycles in aquatic environments. Limnol. Oceanogr., 551–556.

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and Nutrient Cycles in the Sea. Bioscience 49, 781–788. doi:10.2307/1313569.

Zablocki, O., van Zyl, L. J., Kirby, B., and Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. Viruses 9. doi:10.3390/v9110348.

Zablocki, O., van Zyl, L., and Trindade, M. (2018). Biogeography and taxonomic overview of terrestrial hot spring thermophilic phages. Extremophiles 22, 827–837. doi:10.1007/s00792-018-1052-5.

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST: A Fast Phage Search Tool. Nucleic Acids Res. 39, 347–352. doi:10.1093/nar/gkr485.

Zhou, Y., Lin, J., Li, N., Hu, Z., and Deng, F. (2013). Characterization and genomic analysis of a plaque purified strain of cyanophage PP. Virol. Sin. 28, 272–279. doi:10.1007/s12250-013-3363-0.

PUBLICATIONS

frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 September 2018
doi: 10.3389/fmicb.2018.02039

Active Crossfire Between Cyanobacteria and Cyanophages in Phototrophic Mat Communities Within Hot Springs

*Sergio Guajardo-Leiva[1], Carlos Pedrós-Alió[2], Oscar Salgado[1], Fabián Pinto[1] and Beatriz Díez[1,3]**

[1] Department of Molecular Genetics and Microbiology, Pontificia Universidad Católica de Chile, Santiago, Chile, [2] Programa de Biología de Sistemas, Centro Nacional de Biotecnología – Consejo Superior de Investigaciones Científicas, Madrid, Spain, [3] Center for Climate and Resilience Research, Santiago, Chile

Cyanophages are viruses with a wide distribution in aquatic ecosystems, that specifically infect Cyanobacteria. These viruses can be readily isolated from marine and fresh waters environments; however, their presence in cosmopolitan thermophilic phototrophic mats remains largely unknown. This study investigates the morphological diversity (TEM), taxonomic composition (metagenomics), and active infectivity (metatranscriptomics) of viral communities over a thermal gradient in hot spring phototrophic mats from Northern Patagonia (Chile). The mats were dominated (up to 53%) by cosmopolitan thermophilic filamentous true-branching cyanobacteria from the genus *Mastigocladus*, the associated viral community was predominantly composed of Caudovirales (70%), with most of the active infections driven by cyanophages (up to 90% of Caudovirales transcripts). Metagenomic assembly lead to the first full genome description of a T7-like Thermophilic Cyanophage recovered from a hot spring (Porcelana Hot Spring, Chile), with a temperature of 58°C (TC-CHP58). This could potentially represent a world-wide thermophilic lineage of podoviruses that infect cyanobacteria. In the hot spring, TC-CHP58 was active over a temperature gradient from 48 to 66°C, showing a high population variability represented by 1979 single nucleotide variants (SNVs). TC-CHP58 was associated to the *Mastigocladus* spp. by CRISPR spacers. Marked differences in metagenomic CRISPR *loci* number and spacers diversity, as well as SNVs, in the TC-CHP58 proto-spacers at different temperatures, reinforce the theory of co-evolution between natural virus populations and cyanobacterial hosts. Considering the importance of cyanobacteria in hot spring biogeochemical cycles, the description of this new cyanopodovirus lineage may have global implications for the functioning of these extreme ecosystems.

Keywords: hot-springs, cyanophages, phototrophic microbial mat, CRISPR, thermophilic cyanobacteria

OPEN ACCESS

Edited by:
Rui Zhang,
Xiamen University, China

Reviewed by:
Purificacion Lopez-Garcia,
Centre National de la Recherche Scientifique (CNRS), France
Yongle Xu,
Shandong University, China

***Correspondence:**
Beatriz Díez
bdiez@bio.puc.cl

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

Received: *10 April 2018*
Accepted: *13 August 2018*
Published: *03 September 2018*

Citation:
Guajardo-Leiva S, Pedrós-Alió C, Salgado O, Pinto F and Díez B (2018) Active Crossfire Between Cyanobacteria and Cyanophages in Phototrophic Mat Communities Within Hot Springs. Front. Microbiol. 9:2039. doi: 10.3389/fmicb.2018.02039

INTRODUCTION

Hot springs host microbial communities dominated by a limited variety of microorganisms that form well-defined mats (Uldahl and Peng, 2013; Inskeep et al., 2013). Frequently, the uppermost layer of the mat is composed of photoautotrophs; such as oxygenic phototrophic cyanobacteria, including the unicellular cyanobacterium *Synechococcus* spp. (Steunou et al., 2006, 2008;

Bhaya et al., 2007; Klatt et al., 2011), the filamentous non-heterocystous *Oscillatoria* spp., the filamentous heterocystous *Mastigocladus* spp. (Stewart, 1970; Miller et al., 2006; Mackenzie et al., 2013; Alcamán et al., 2015), as well as filamentous anoxygenic phototrophs (FAPs), such as *Roseiflexus* sp. and *Chloroflexus* sp. (Van der Meer et al., 2010; Klatt et al., 2011; Liu et al., 2011). These primary producers interact with heterotrophic prokaryotes through element and energy cycling (Klatt et al., 2013). Heterocystous cyanobacteria are a key component in hot springs, since these systems are commonly N-limited due to the rapid assimilation and turnover of inorganic nitrogen forms (Alcamán et al., 2015; Lin et al., 2015). Thus, N_2-fixation by cyanobacteria is identified to be a key biological process in neutral hot spring microbial mats (Alcamán et al., 2015).

These simplified but highly cooperative communities have been historically used as models for understanding the composition, structure, and function of microbial consortia (Klatt et al., 2011; Inskeep et al., 2013). The role of a variety of abiotic factors, such as pH, sulfide concentration, and temperature, in determining microbial assemblages and life cycles in these ecosystems have been investigated (Cole et al., 2013; Inskeep et al., 2013). However, there is a lack of investigation into biotic factors, such as viruses, on thermophilic photoautotrophic mats, with existing studies only reporting short or partial viral sequences (Heidelberg et al., 2009; Davison et al., 2016). Currently, viral communities from thermal mats have been characterized through indirect approaches, indicating the hypothetical presence of viruses (Heidelberg et al., 2009; Davison et al., 2016). Heidelberg et al. (2009) used CRISPR spacer sequences extracted from the genomes of two thermophilic *Synechococcus* isolates, from a phototrophic mat in Octopus Spring. Subsequently, they searched for viral contigs from previously published water metaviromes from the Octopus and Bear Paw Springs in Yellowstone National Park (United States) (Schoenfeld et al., 2008). Furthermore, Davison et al. used CRISPR spacers and nucleotide motive frequencies to link viral contigs to known hosts using a metavirome obtained by Multiple Displacement Amplification (MDA) of VLPs from a mat in Octopus Spring (Davison et al., 2016), as well as reference genomes from dominant species (*Synechococcus* sp., *Roseiflexus* sp., and *Chloroflexus* sp.) previously described in the same microbial mat. A key finding from these studies was the link between viruses and their hosts, indicating their co-evolution and an effective "arms race" within hot spring phototrophic mats.

Unlike thermophilic mat studies, most viral investigation carried out in hot springs occur within the source waters (Rachel et al., 2002; Yu et al., 2006; Schoenfeld et al., 2008; Bolduc et al., 2012, 2015; Zablocki et al., 2017). In these waters, virus abundances range between 10^4 and 10^9 virus like particles (VLPs) mL^{-1} (Breitbart et al., 2004; Schoenfeld et al., 2008; Redder et al., 2009). They play an important role in both the structuring of host populations and as drivers of organic and inorganic nutrient recycling (Breitbart et al., 2004). The majority of the viruses were dsDNA, with new and complex viral morphotypes, distinct to the typical head and tail morphologies (Rachel et al., 2002; Prangishvili and Garrett, 2004; Schoenfeld et al., 2008; Redder et al., 2009; Pawlowski et al., 2014). Furthermore, the few metaviromes obtained in thermal waters indicate that natural thermophilic virus communities differ from those obtained in culture, given that there was only a 20–50% similarity between the sequences obtained and those in the databases (Pride and Schoenfeld, 2008; Schoenfeld et al., 2008; Diemer and Stedman, 2012; Bolduc et al., 2015). Thus far, the genomes that have been isolated and sequenced from thermophilic viruses (57 genomes, of which 37 infected archaea and 20 infected Bacteria) generally yielded few significant matches to sequences in public databases (Uldahl and Peng, 2013). More recently, a water metaviromic study from Brandvlei hot spring (BHS), South Africa (Zablocki et al., 2017) reported the presence of two partial genomes (10 kb and 27 kb), the first related to Podoviridae and the second to lambda-like Siphoviridae families. Both Caudovirales genomes did not have a confirmed host, but the presence of green microbial mat-patches around the contours of the hot spring, implied that filamentous Cyanobacteria and unclassified *Gemmata* species were the potential hosts, respectively. The last, based on the proximity of some viral predicted proteins with bacteria from well characterized microbial mats present in a nearby hot spring (Tekere et al., 2011; Jonker et al., 2013).

Given the lack of knowledge of viral communities within hot spring phototrophic microbial mats, the present study used the mats of Porcelana hot spring (Northern Patagonia, Chile), as a pH neutral model, to better understand the associated thermophilic viral communities within these mats. This pristine spring is covered by microbial mats that grow along a thermal gradient between 70 and 46°C, dominated by bacterial phototrophs, such as filamentous cyanobacteria from the genus *Mastigocladus* (Mackenzie et al., 2013; Alcamán et al., 2015). This is the dominant and most active cyanobacterial genus in the Porcelana mat environment, carrying out important biological processes such as carbon- and N_2-fixation (Alcamán et al., 2015, 2017). Thus, this study proposes that the mats in Porcelana hot spring are dominated by viral communities of the Order Caudovirales, which is able to infect Cyanobacteria, preferably *Mastigocladus* spp.

The viral diversity in Porcelana was determined through the detection of viral signals in microbial mat omics data, and by TEM along the thermal gradient. The results demonstrate that the viral community was dominated by Caudovirales, which actively infect Cyanobacteria. Furthermore, the first complete genome description of a thermophilic cyanobacterial T7-like podovirus, Thermophilic Cyanophage Chile Porcelana 58°C (from now on TC-CHP58) is realized. The host is the dominant phototroph *Mastigocladus* spp, based on CRISPR spacers. Finally, the presence of different populations of this new podovirus are identified through single nucleotide variants (SNVs) analyses, and the co-evolution of *Mastigocladus* spp. and particular populations of TC-CHP58 at different temperatures is described through association of specific SNVs to different CRISPR spacers.

MATERIALS AND METHODS

Sampling Site

Porcelana hot spring is located in Chilean Patagonia (42° 27′ 29.1′′S – 72° 27′ 39.3′′W). It has a neutral pH range between 7.1 and 6.8 and temperatures ranging from 70 to 46°C, when sampled on March 2013. Phototrophic microbial mats growing at 66, 58, and 48°C were sampled using a cork borer of 7 mm diameter. Cores of 1 cm thick were collected in triplicate at noon (12:00 PM), transported in liquid nitrogen and kept at −80°C until DNA and RNA extraction.

Transmission Electron Microscopy

Five liters of interstitial fluid was squeezed using 150 μm sterilized polyester net SEFAR PET 1000 (Sefar, Heiden, Switzerland) and filtered through 0.8 μm pore-size polycarbonate filters (Isopore ATTP, 47 mm diameter, Millipore, Millford, MA, United States) and 0.2 μm pore-size (Isopore GTTP, 47 mm diameter, Millipore) using a Swinex filter holder (Millipore). Particles in the 0.2 μm filtrate were concentrated to a final volume of approximately 35 ml using a tangential-flow filtration cartridge (Vivaflow 200, 30 kDa pore size, Vivascience, Lincoln, United Kingdom). Viral concentrates (15 μL) were spotted onto Carbon Type-B, 200 mesh, Copper microscopy grids (Ted Pella, Redding, California, United States), stained with 1% uranyl acetate and imaged on an FEI Tecnai T12 electron microscope at 80 kV (FEI Corporate, Hillsboro, OR, United States) with attached Megaview G2 CCD camera (Olympus SIS, Münster, Germany). Imaging analysis was done at the Advanced Microscopy Unit, School of Biological Sciences at Pontificia Universidad Católica de Chile (Santiago, Chile).

Nucleic Acid Extractions and High Throughput Sequencing

Nucleic acids (DNA and RNA) were extracted as previously described (Alcamán et al., 2015). For RNA, Trizol (Invitrogen, Carlsbad, CA, United States) was added to the mat sample, and homogenized by bead beating, two pulses of 20 s. Quality and quantity of the extracted nucleic acids were checked and kept at −80°C.

Samples were sequenced by Illumina Hi-seq technology (Research and Testing Laboratory, Texas, United States). Briefly, for metagenomes, DNA was fragmented using NEBNext dsFragmentase (New England Biolabs, Ipswich, MA, United States), followed by DNA clean up using column purification, and a NEBUltra DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, United States) was used for library construction.

For metatranscriptomes, DNase treated total RNA was cleaned up of rRNA by a Ribo-Zero rRNA Removal Kit Bacteria (Illumina, San Diego, CA, United States), followed by purification using an Agencourt RNAClean XP Kit (Beckman Coulter, Indianapolis, IN, United States), and a NEXTflex^{TM} Illumina Small RNA Sequencing Kit v3 (Bio Scientific, Austin, TX, United States) was used for library construction.

For quality filtering, the following filters were applied using Cutadapt (Martin, 2011), leaving only mappable sequences longer than 30 bp (-m 30), with a 3′ end trimming for bases with a quality below 28 (-q 28), a hard clipping of the first five leftmost bases (-u 5), and finally a perfect match of at least 10 bp (-O 10) against the standard Illumina adaptor. Finally, the removal of sequences representing simple repetitions that are usually due to sequencing errors was applied using PRINSEQ (Schmieder and Edwards, 2011) DUST threshold 7 (-lc_method dust, -lc_threshold 7). Details of the number of sequences obtained are shown in **Supplementary Table S1**.

Identification of rRNA-Like Sequences and Viral Mining From Metagenomes and Metatranscriptomes

Metagenomic Illumina TAGs (miTAGs) (Logares et al., 2014) that are small subunit (SSU) 16S and 18S rRNA gene sequences in the metagenomes were identified and annotated using the Ribopicker tool (Schmieder et al., 2012) with the Silva 123 SSU database (Quast et al., 2013).

For viral mining, bacterial, archaeal and eukaryotic sequences were removed through end-to-end mapping, allowing a 5% of mismatch (-N 1 -L 20) against the NCBI non-redundant (NR) database (Nov-2015) using bowtie2 (Langmead and Salzberg, 2012). Viral sequences were then recruited against modified NCBI RefSeq (Release 75) viral proteins, where only amino acid sequences from viruses that do not infect animals (NAV) were considered to build the database, using the UBLAST algorithm (-strand both -accel 0.9) through the USEARCH sequence analysis tool (Edgar, 2010). Recruitment was made for sequences with over 65% of coverage and an E-value $< 1 \times 10^{-3}$ (-query_cov 0.65 -evalue 1e-3). For taxonomic assignment, recruited sequences were aligned against the NAV database using BLASTX (Camacho et al., 2009) and parsed using the lowest common ancestor algorithm trough MEGAN 6 (Huson et al., 2016) (LCA score = 30). The latter displays a graphical representation of abundance for each taxonomic group identified at the family and species levels. Species classification of viral reads, was used to infer the phyla of the putative hosts based on viral RefSeq host information or through a manual search of the publication associated with each viral genome.

To extract putative viral genomes, all metagenomes (48, 58, and 66°C) were assembled using De Bruijn graphs as implemented in the Spades assembler (Bankevich et al., 2012), followed by gene prediction using Prodigal software (Hyatt et al., 2010) and the recovery of circular contigs over 5 kb using a Python script (Crits-Christoph et al., 2016). Only sequences over 5 kb were used in the subsequent analysis because all dsDNA viruses in the databases have genomes over that size. A homology search of the viral predicted proteins by Prokka (Seemann, 2014) was done using BLASTX against the NAV protein database and NCBI nr as described before. Additionally, all contigs over 5 kb were analyzed using VirSorter (Roux et al., 2015a) against the virome database option.

To quantify the abundance and activity of the retrieved viral genome, reads recruitment from each metagenome and

metatranscriptome was performed using BWA-MEM (-M), resulting SAM file was parsed by BBmap pileup script (Bushnell B.)[1].

Phylogenetic Analysis

The protein inferred sequences of DNA polymerase and major capsid were aligned by Muscle (Edgar, 2004) and MAFFT (Katoh et al., 2002), respectively, using the amino acid substitution model determined by ProtTest 3 (Blosum62+G+F) (Darriba et al., 2011) and modelFinder (LG+F+G4), respectively. The Bayesian Markov chain Monte Carlo method was implemented with MrBayes 3.6 (Ronquist et al., 2012) and MCMC results were summarized with Tracer 1.6[2]. MrBayes was run using two independent runs, four chains, 1,500,000 generations and a sampling frequency of 100 with a burn-in value of 33% until the standard deviations of split frequencies remained below 0.01.

The maximum likelihood method was implemented with IQtree (-bb 10000 -nm 10000 -bcor 1 -numstop 1000) (Trifinopoulos et al., 2016) using 100 standard bootstrap and 10,000 ultrafast bootstrap to evaluate branch supports. The details of the sequences used for phylogenetic analyses are listed in **Supplementary Table S2**.

CRISPR/Cas Virotopes

Assemblies for each temperature, were taxonomically grouped (bins) using the Expectation–Maximization (EM) algorithm implemented in MaxBin 2.0 (Wu et al., 2016). In order to asses the completeness and contamination of each bin, CheckM (Parks et al., 2015) analyses were performed. Finally, the closest genome of each bin was searched using the Tetra Correlation Search (TCS) analysis implemented in Jspecies tool (Richter et al., 2016) with selection criteria of Z score greater than 0.999 and ANI over 95% (Konstantinidis et al., 2017).

CRISPR/Cas *loci* were identified in contigs assigned to *Mastigocladus* spp. from 48, 58, and 66°C assembled metagenomes using CRISPRFinder tool (Grissa et al., 2007). To quantify the activity of the CRISPR *loci*, reads recruitment from metatranscriptomes for the same temperatures was performed using BWA-MEM (-M), and the resulting SAM file was parsed by BBmap pileup script (Bushnell B.) (see footnote 1) and normalized by total number of reads and length of each *loci*.

Spacers from CRISPR containing contigs were mapped to viral contigs using bowtie2 (Langmead and Salzberg, 2012) parameters (-end-to-end -very sensitive -N 1). Mapped spacers were manually annotated to the viral predicted proteins in viral contig.

Single Nucleotide Variants (SNVs)

To call variants occurring in TC-CHP58 populations at the three different metagenome temperatures, LoFreq method (Wilm et al., 2012) was used. SNVs frequencies were quantified in ORFs from TC-CHP58 genome using Bedtools suite (Quinlan and Hall, 2010). The alleles of SNVs present in proto-spacers were visualized in IGV tools for each virotope at each temperature.

[1] sourceforge.net/projects/bbmap/

[2] http://beast.bio.ed.ac.uk/Tracer

RESULTS

Morphological and Genetic Composition of VLPs

Transmission electron microscopy (TEM) was applied to identify the VLPs present in the interstitial fluid from microbial mats in Porcelana hot spring. Caudovirus-like particles belonging to Myoviridae, Podoviridae and Siphoviridae families, typically infecting bacteria (**Figures 1A–G**) were identified. Additionally, filamentous and rod shaped VLPs were detected, that could be associated with Lipothrixviridae and Clavaviridae families, usually infecting archaea (**Figures 1H–K**). Viral read counts ranged between 0.47 and 0.78% of the total metagenome reads, and between 0.35 and 3.71% in the metatranscriptomes (**Supplementary Table S1**). At all temperatures, viral metagenomic sequences (**Figure 2**) revealed the dominance of the Order Caudovirales, followed by the Order Megavirales, with ∼70% and ∼23% of the total viral reads, respectively. Metatranscriptomic analysis results (**Figure 2**) showed a slightly different pattern, with a reduction in Caudovirales with increasing temperature (from ∼78% at 48°C to ∼57% at 66°C), whereas Megavirales did the opposite (from ∼7% at 48°C to ∼36% at 66°C).

In the metagenomes, Siphoviridae was the most abundant family of Caudovirales, with maximum abundance at 48°C. Myoviridae members were also well represented with a maximum of ∼31% at 58°C and a minimum (∼25%) at 48°C. Meanwhile, Podoviridae accounted for just ∼8% at all temperatures (**Figure 2**). In metatranscriptomes, Siphoviridae increased sixfold with temperature, while Podoviridae and Myoviridae decreased with temperature (fivefold and twofold, respectively).

The Megavirales order was also present, however, at a lower abundance compared to Caudovirales. Megavirales were represented by Phycodnaviridae (∼13%), Mimiviridae (∼8%), and Marseilleviridae (∼2%) families, remaining constant through all temperatures. Metatranscriptomics showed an increase in abundance of these three virus families with temperature.

Caudovirales Host Assignments

Porcelana mat communities based on miTAGs were dominated by bacteria (∼96%), with low abundances of eukarya (∼3%) and archaea (∼1%) (**Supplementary Table S1**). At the phylum level (**Figure 3A**), bacterial communities were mostly composed of Cyanobacteria oxygenic phototrophs (33, 53, and 21% of total rRNA SSU sequences at 48, 58, and 66°C, respectively) and Chloroflexi anoxygenic phototrophs (higher than Cyanobacteria only at 66°C, with 35% of total rRNA SSU sequences). Other representative members of the community were Proteobacteria (5–11%), Deinococcus–Thermus (2–7%), Firmicutes (1–17%), and Bacteroidetes (4–8%) (**Figure 3A**).

The host assignment, based on taxonomy from viral reads of the most representative Caudovirales (**Figure 3B**), showed that viruses putatively infected members of the bacterial phyla Proteobacteria, Cyanobacteria, Actinobacteria, and Firmicutes. Metagenomic data showed that increases in temperature led to an increase in viruses from Actinobacteria and Firmicutes.

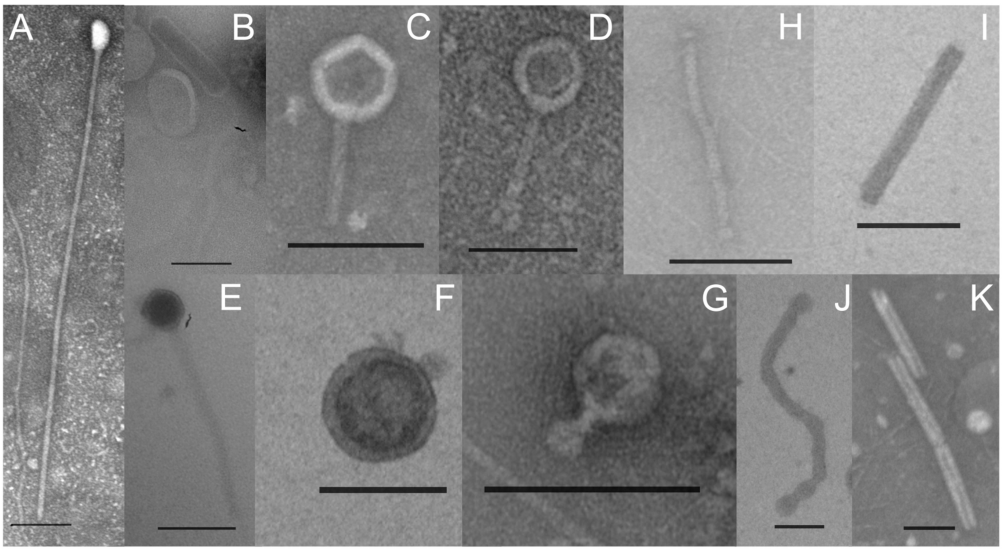


FIGURE 1 | Transmission electronic micrographs of VLPs obtained from the interstitial fluid of phototrophic microbial mats growing between 62°C and 42°C in Porcelana hot spring. Scale bar: 100 nm. **(A–G)** Caudovirus-like particles belonging to Myoviridae, Podoviridae, and Siphoviridae families. **(H–K)** Filamentous and rod shaped VLPs that could be associated with Lipothrixviridae and Clavaviridae families.

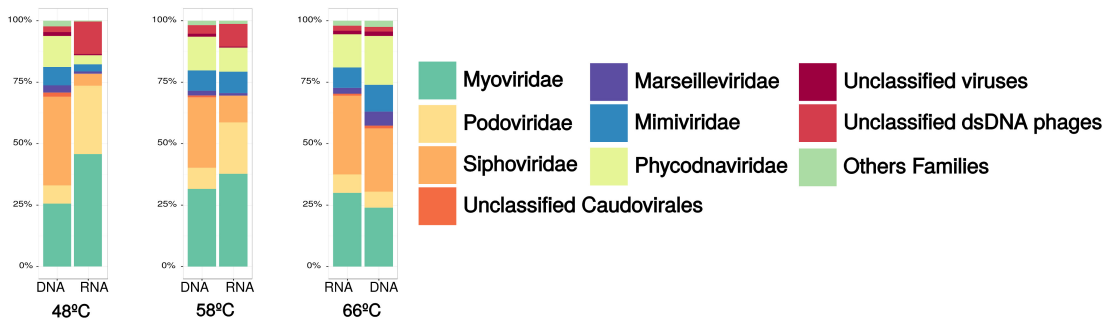


FIGURE 2 | Relative abundances of viral Families in microbial mats from Porcelana hot spring; standardized by the total number metagenomic reads (DNA), and metatranscriptome (RNA) from each temperature samples.

Additionally, an increase in Cyanobacteria viruses was observed at 58°C. Viruses from Proteobacteria, Actinobacteria, and Firmicutes were represented by the three Caudovirales families, while viruses from Cyanobacteria were represented by Podoviridae and Myoviridae families only (**Supplementary Table S3**), where cyanopodovirus and cyanomyovirus reads increase from 31 to 50% at 48°C and from 30 to 45% at 58°C, then decrease to 23 to 28% at 66°C, respectively.

Metatranscriptomic sequences from Caudovirales potentially infecting Cyanobacteria, were predominant at 48°C and 58°C, with ∼90% and ∼74% of the total viral sequences, respectively. However, cyanophage transcripts abruptly decrease at 66°C. Cyanophages were exclusively related to the Myoviridae and Podoviridae families (**Supplementary Table S3**). Reads associated with cyanopodoviruses and cyanomyoviruses gradually decreased with temperature; between 48 and 58°C, virus reads declined from 95% and 96% to 84% and 89%, respectively. On the other hand, at 66°C a more severe decline was observed, to 15% and 20%, respectively. Conversely, with the reduced representation of Cyanobacteria at 66°C, other caudovirales transcripts increased, including those that infect Proteobacteria (∼31%), Firmicutes (∼30%), and Actinobacteria (∼23%).

Thermophilic Cyanophage Genome Recovery

The metagenome assembly recovered 3,912; 2,697; and 2,758 contigs, at 48°C, 58°C, and 66°C, respectively. A script search (Crits-Christoph et al., 2016) resulted in 11 circular contigs, possibly indicating complete genomes. Subsequent BLASTP analysis (Camacho et al., 2009) of predicted proteins indicated that only one circular contig had viral hallmark genes, meanwhile nine contigs had genes associated with bacterial mobile genetic elements and one contig remain completely unknown. These hallmark genes are shared by many viruses but are absents

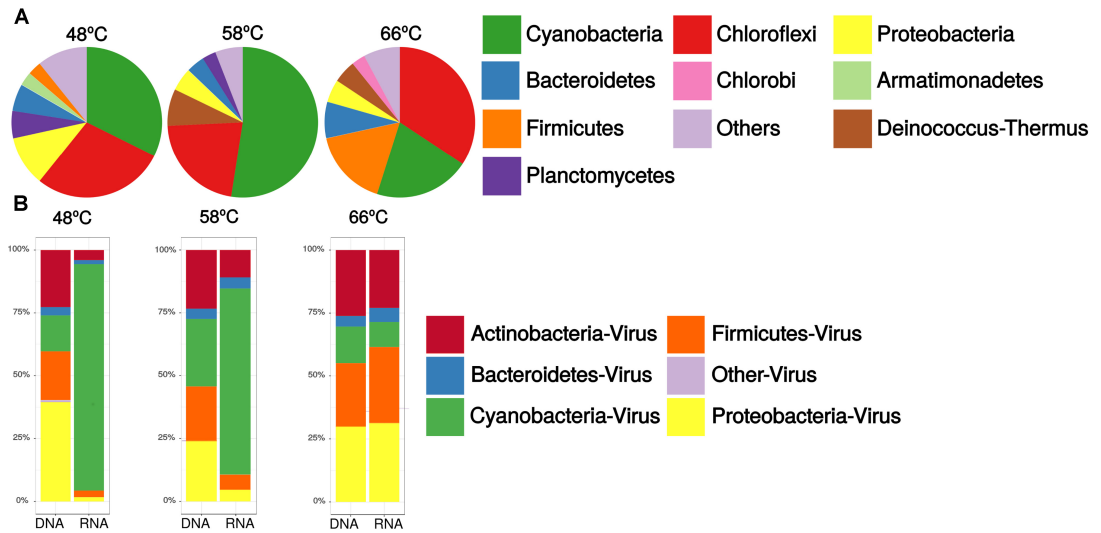


FIGURE 3 | Relative abundances of **(A)** Bacterial community, to Phylum level, in the microbial mats obtained from 16S miTAGs, standardized by the total number of metagenomic reads (DNA) for each temperature sample, and **(B)** Caudovirales community at the host Phylum level, obtained from shotgun sequences in metagenomes (DNA) and metatranscriptomes (RNA), standardized by the total number of reads from each temperature sample.

from cellular genomes (Koonin et al., 2006). VirSorter tool analysis (Roux et al., 2015a) confirmed these results, obtaining the same complete putative viral contig from the 58°C assembly, 40,740 bp long and 43.9% of GC content. This contig, TC-CHP58 (**Figure 4A**), was associated with a Cyanobacterial host. TC-CHP58 was present (reads recruitment) over all temperatures in Porcelana hot spring (**Figure 5** and **Supplementary Figure S1**). At 66°C, TC-CHP58 was sevenfold more abundant than their putative host (measured as *Mastigocladus* RUBISCO gene abundance); at 48°C, the virus-host ratio was 1:1, and at 58°C the host was fourfold more abundant than TC-CHP58. Metatrancriptomic reads also show that TC-CHP58 was active over all temperatures (**Figure 5** and **Supplementary Figure S2**), but with lower transcription levels than the putative host (measured as *Mastigocladus* RUBISCO gene activity), ranging between 80- and 8-fold lower (**Supplementary Table S4**). TC-CHP58 viral DNA:RNA ratio indicated similar proportions (2.4) at 58°C, least similar (552.9) at 66°C; while at 48°C the ratio was 10.4 (**Supplementary Table S4**).

Genomic Features and Organization of TC-CHP58

Complete protein prediction and annotation of TC-CHP58 using Prokka (Seemann, 2014) and BLASTP revealed 39 putative ORFs, 10 of which were viral core proteins (i.e., capsid and tail-related proteins, DNA polymerase, Terminase, etc.), 22 had no significant similarities in NCBI nr database, and 4 were present in the database but with unknown function (**Table 1**).

Blast analysis of the viral genes in TC-CHP58, revealed 25 to 48% identity (amino acidic level) with proteins from Cyanophage PP, PF-WMP3 and Anabaena phage A-4L, that infect freshwater filamentous Cyanobacteria such as *Phormidium*, *Plectonema,* and *Anabaena* (**Table 1**). At the nucleotide level, there was almost no similarity to any known sequence except for a short segment of 40 nucleotides, which showed 93% similarity to a Portal protein gene sequence of *Plectonema* and P*hormidium* cyanopodoviruses (Cyanophage PP; NC_022751 and PF-WMP3; NC_009551).

Gene prediction by Prodigal indicated that the TC-CHP58 genome might be structured into two clusters, based on the transcriptional direction and putative gene functions (**Figure 4A**). The predicted ORFs (**Table 1**) in the sense strand encode proteins involved in DNA replication and modification, such as DNA polymerase and DNA primase/helicase. Conversely, the ORFs in the antisense strand (**Table 1**) encode proteins necessary for virion assembly, such as major capsid protein (MCP), tail fiber proteins, internal protein/peptidase, tail tubular proteins, scaffold protein, and portal protein. Moreover, two ORFs in the antisense strand had the best hits to the cyanobacterial hypothetical proteins found in the filamentous cyanobacterium *Fischerella* (WP_026731322. 1) and the unicellular *Gloeobacter* (WP_023172199.1).

Additionally, VIRFAM (Lopes et al., 2014) was used to classify TC-CHP58 according to their neck organization (**Supplementary Figure S3**), being assigned to the Podoviridae Type 3 category with neck structural organization similar to the Enterobacteria phage P22 (Lopes et al., 2014). Hierarchical clustering of neck proteins grouped TC-CHP58 together with the freshwater cyanophages Pf-WMP3 and Pf-WMP4, separating them from marine cyanophages such as P60 and Syn5.

Even when a large number of viral reads were assigned to cyanophages of Myoviridae family, it was not possible to recover any genome of this type. Most of the Myoviridae related contigs only had non-structural genes or hypothetical proteins of unknown function which align with proteins of known

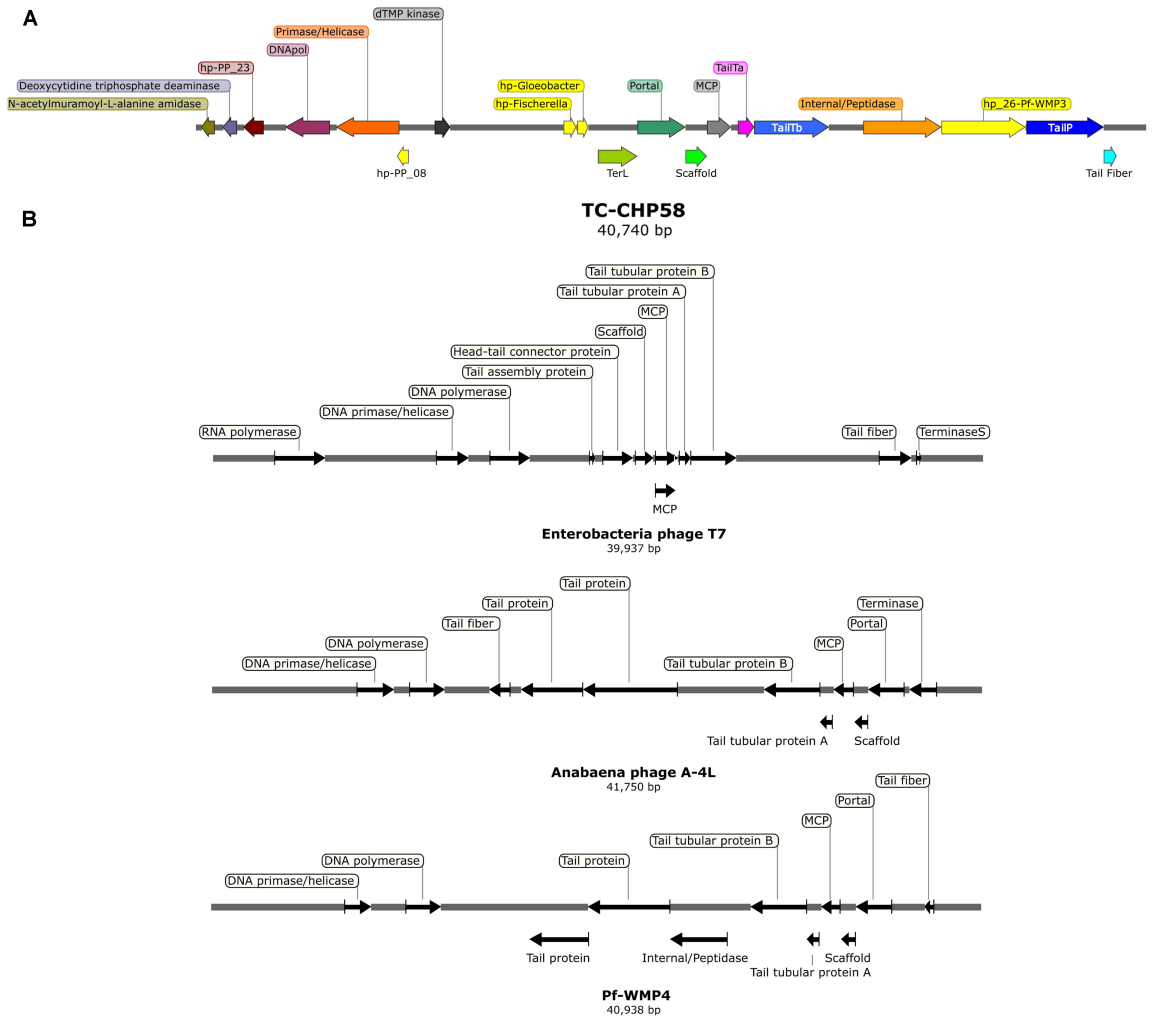


FIGURE 4 | (A) Genomic organization of the thermophilic cyanophage CHP58 (TC-CHP58). Arrows indicate the size, position, and orientation of annotated ORFs, with predicted functions or homologs (e.g., DNApol, DNA polymerase; TailY a/b, tail tubular protein a/b; MCP, major capsid protein; TerL, large terminase subunit; TailP, Tail protein; hp-PP_08/23, homologous to hypothetical proteins 08/23 from cyanophage PP; hp-*Fischerella/Gloeobacter*, homologous to hypothetical proteins in *Fischerella* sp. PCC 9605/*Gloeobacter kilaueensis*). **(B)** Genomic organization of Enterobacteria phage T7, Anabaena phage A4L and Pf-WMP4. Arrows indicate the size, position, and orientation of viral core ORFs.

cyanomyoviruses. Here, the absence of hallmark genes from Cyanobacteria related viruses makes their accurate classification as cyanomyoviruses impossible.

Phylogenetic Analysis of Phage TC-CHP58

To investigate the relationship of the phage TC-CHP58 within the Podoviridae family, the DNApol gene was selected for comparison, using published viral genomes. The analysis included representatives of Picovirinae and Autographivirinae subfamilies, plus all the available DNApol genes from known freshwater podoviruses (Pf-WMP3, PP, Pf-WMP4 and A-4L) infecting filamentous heterocystous cyanobacteria from the order Nostocales and non-heterocystous from order Oscillatoriales, plus those infecting marine *Synechococcus* spp. and *Prochlorococcus* spp. The DNApol tree (**Figure 6**) showed the phage TC-CHP58 as part of a monophyletic clade with all cyanopodoviruses described as infecting freshwater filamentous cyanobacteria, and more distantly, with the marine cyanopodovirus clade that infects *Synechococcus* spp. and *Prochlorococcus* spp. Both cyanophage subgroups are closely related with podoviruses from the Autographivirinae subfamily, which includes all T7 relatives. Furthermore, the phylogeny of the MCP was constructed for freshwater and marine representatives of the Autographivirinae subfamily. The available MCP gene from BHS3 Cyanophage partial genome, that is the only known thermophilic representative within the Podoviridae family, was also included (Zablocki et al., 2017). The MCP tree (**Supplementary Figure S4**) showed similar results to the DNApol tree (**Figure 6**), with a monophyletic origin for all freshwater cyanophages infecting

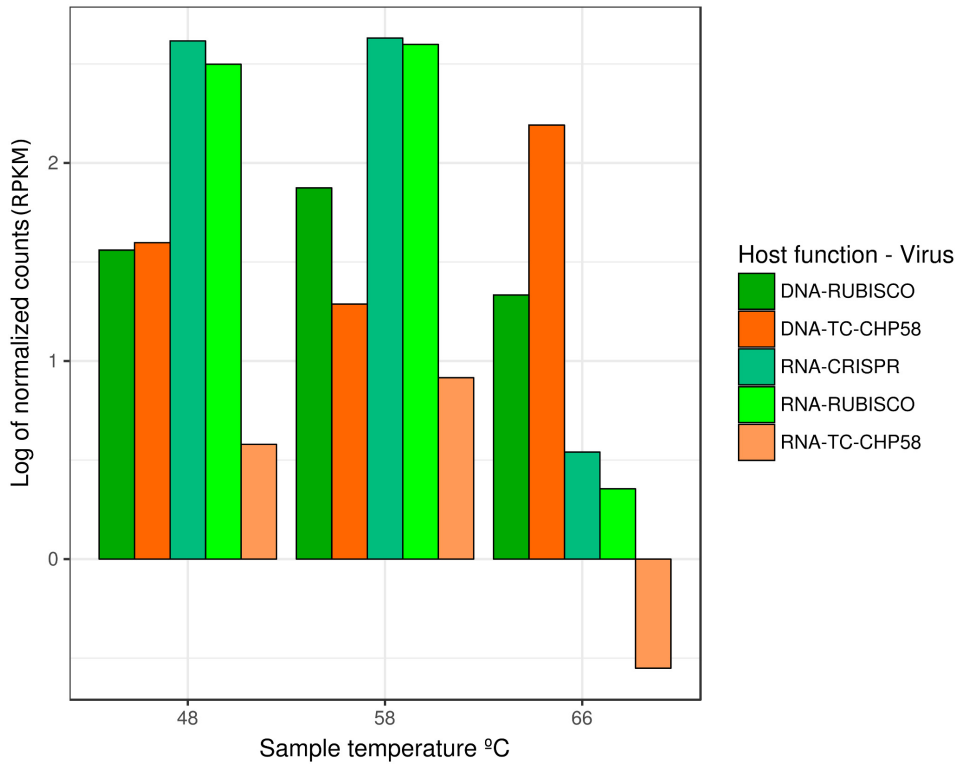


FIGURE 5 | Relative abundance and transcriptomic expression of *Mastigocladus* spp. CRISPR systems and TC-CHP58. Abundance and expression of *Mastigocladus* RUBISCO was used as reference of the cyanobacterial presence and metabolic activity. Only specific CRISPR *loci* with proto-spacers in TC-CHP58 were fully quantified for each temperature. For improved visualization, counts are represented as Log of reads per kilobase million (RPKM).

TABLE 1 | Blastp analysis of predicted CDS from TC-CHP58 of known function against NCBI RefSeq (Release 75) and NR databases.

Query sequence ID	Subject sequence ID	Identity %	E-value	Bit Score
TC-CHP58_sequence	KF598865.1| [Cyanophage PP]	93%	0.2	54.7
TC-CHP58_CDS1	YP_009042789.1| DNA polymerase [Anabaena phage A-4L]	29.03	4.00E-60	223
TC-CHP58_CDS3	YP_009042786.1| DNA primase/helicase [Anabaena phage A-4L]	25.21	2.00E-28	131
TC-CHP58_CDS4	YP_008766966.1| hypothetical protein PP_08 [Cyanophage PP]	29.7	0.0006	47
TC-CHP58_CDS7	WP_026824764.1| dTMP kinase [*Exiguobacterium marinum*]	30.61	4.00E-22	99.4
TC-CHP58_CDS13	WP_026731322.1| hypothetical protein [*Fischerella* sp. PCC 9605]	34.55	7.00E-12	69.3
TC-CHP58_CDS14	WP_023172199.1| hypothetical protein [*Gloeobacter kilaueensis*]	42.31	4E-05	48.1
TC-CHP58_CDS15	YP_008766995.1| terminase [Cyanophage PP]	44.39	2.00E-150	456
TC-CHP58_CDS16	YP_001285799.1| portal protein [Phormidium phage Pf-WMP3]	42.96	0	551
TC-CHP58_CDS17	YP_009042804.1| scaffold protein [Anabaena phage A-4L]	30.69	1E-07	60.8
TC-CHP58_CDS18	YP_008766991.1| capsid protein [Cyanophage PP]	48.14	4.00E-109	335
TC-CHP58_CDS20	YP_009042802.1| tail tubular protein A [Anabaena phage A-4L]	29.52	3.00E-28	116
TC-CHP58_CDS21	YP_001285795.1| tail tubular protein B [Phormidium phage Pf-WMP3]	36.49	0	630
TC-CHP58_CDS24	YP_009042798.1| internal protein [Anabaena phage A-4L]	29.86	5.00E-41	174
TC-CHP58_CDS25	YP_001285791.1| PfWMP3_26 [Phormidium phage Pf-WMP3]	28.26	5.00E-23	117
TC-CHP58_CDS26	YP_009042796.1| tail protein [Anabaena phage A-4L]	24.8	4.00E-89	322
TC-CHP58_CDS32	WP_038085449.1| *N*-acetylmuramoyl-L-alanine amidase [*Tolypothrix bouteillei*]	46.29	3.00E-46	160
TC-CHP58_CDS35	WP_043587103.1| deoxycytidine triphosphate deaminase [*Diplosphaera colitermitum*]	44.9	2.00E-48	167
TC-CHP58_CDS37	YP_008766981.1| hypothetical protein PP_23 [Cyanophage PP]	29.92	1.00E-21	101

filamentous cyanobacteria, emphasizing the division between freshwater and marine cyanobacterial viruses, and their affiliation with T7 phage. The thermophilic representatives of Podoviridae family were located in different branches inside the freshwater clade, with BHS3 more basal than TC-CHP58.

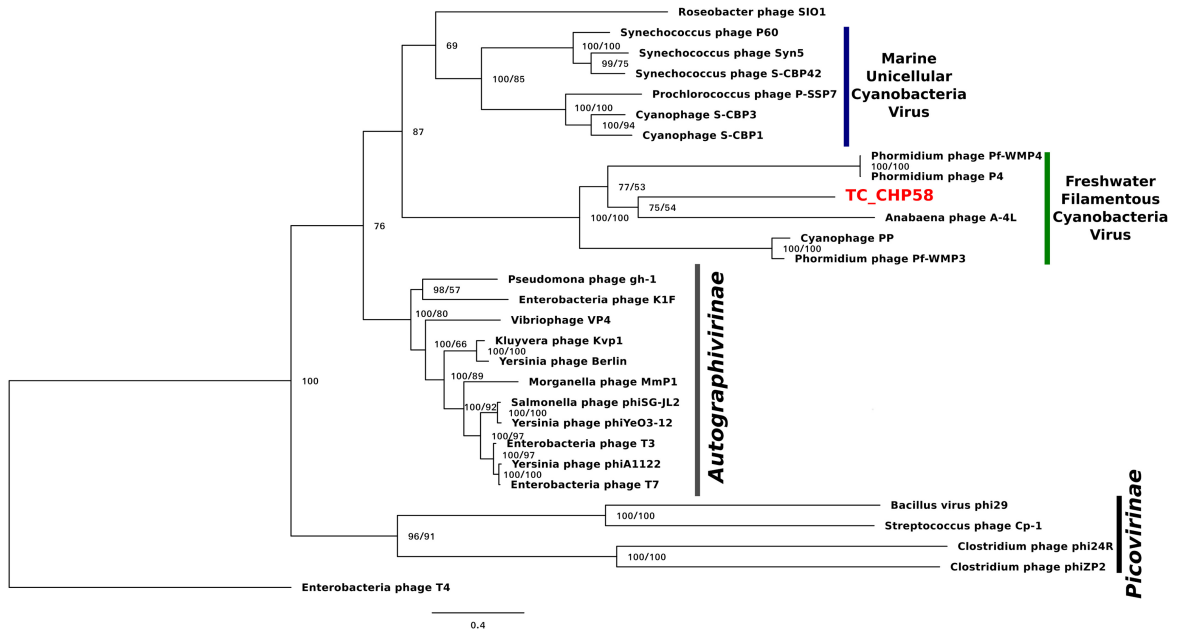


FIGURE 6 | Bayesian inference phylogenetic reconstruction of DNA polymerase I protein of TC-CHP58. Numbers indicate Bayesian posterior probabilities as percentage/ultra-fast bootstrap values. Only UFBoot values over 80 and Bayesian PP over 50 are shown. The sequence characterized in the present study is reported in bold letters. Scale bar: 0.4 amino acid substitutions per site.

CRISPR Arrays on TC-CHP58 Host

Given the high abundance (Mackenzie et al., 2013; Alcamán et al., 2015) and activity (Alcamán et al., 2015) of cyanobacteria, such as *Mastigocladus* spp., in Porcelana hot spring (**Figure 3A**), and in order to confirm the putative host of phage TC-CHP58, CRISPR spacer arrays were identified using the CRISPRFinder tool (Grissa et al., 2007) for seven *Mastigocladus* spp. Contigs, obtained from metagenome assemblies at 48°C, 58°C, and 66°C. Three CRISPR *loci* were common between all temperatures (48_CRISPR_2, 58_CRISPR_5, and 66_CRISPR_2), while four *loci* were specific to higher temperatures (58–66°C) (**Table 2**). In total, the seven CRISPR *loci* contain 562 spacers, of which 25 of them had a proto-spacer sequence in the TC-CHP58 genome (**Table 2**). From the 25 spacers, 19 have a target ORF of known function, such as DNA polymerase, dTMP, portal protein, M23-petidase, tail protein, tail fiber, and deoxycytidine triphosphate deaminase. In general, each CRISPR *loci* contained spacers against different ORFs on TC-CHP58, or even against different locations on the same ORF. For the 25 spacers, searching the nt/nr database, using BLASTN and BLASTX, showed no similarity to any know sequence. Finally, in order to check if CRISPR systems were active, expression of the seven *loci* was directly quantified in the three metatranscriptomes. For all temperatures, slightly lower transcript levels were found compared to the *Mastigocladus* RUBISCO gene (**Figure 5**).

Identifying Single Nucleotide Variants in TC-CHP58 Genome

To assess if mismatches between the CRISPR spacer and proto-spacer sequences in TC-CHP58 genome were concealing potential variations in TC-CHP58 populations, a SNV calling was conducted. For this task LoFreq tool was used, as it is high sensitivity and has low false positive rates, lower as <0.00005% (Wilm et al., 2012) and higher as 8.3% (Huang et al., 2015). This approach, together with the use of sequences with qualities over q28 (whose error probability in the base call is ≤1.58%), allow us to consider these SNVs as real mutations.

A different number of SNVs was found at each temperature. TC-CHP58 showed 1611, 930, and 671 variant sites at 48°C, 58°C, and 66°C, respectively, unevenly distributed throughout the viral genome (**Supplementary Figure S5**). Considering the three metagenomes, a total of 3212 variable sites were present in the TC-CHP58 genome, with 391 SNVs present over all temperatures (**Supplementary Figure S5**). Most of the SNVs (74% on average) were located at coding regions on the TC-CHP58 genome, with variable rates, ranging from 15 to 0 SNVs for each 100 bp (**Supplementary Table S5**) over different ORFs.

A detailed analysis of SNVs in CRISPRs proto-spacer sites revealed the presence of these polymorphisms in 14 of the 25 spacer targets, with 13 mismatches and 4 perfect matches (**Table 2**). The total number of polymorphic sites was 22, with 13 SNVs causing a synonymous substitution and 7 causing a non-synonymous substitution (**Table 2**).

DISCUSSION

The study of viruses from thermophilic phototrophic microbial mat communities remains largely unexplored except for a few cases providing limited information on viral presence

TABLE 2 | CRISPR *loci* at each temperature detailed information and SNVs analysis for TC-CHP58 proto-spacer including alleles frequency and SNV coding effect.

T°C Sample	Virotope Sequence	Viral target	CRISPR *loci*	Proto-spacer Start	Proto-spacer End	Mismatch position	SNV position	Alleles *	Frequency of in CRISPR allele	SNV effect	Codon change
48	ACCTTTCAGACCTAACTC TAAAGTTACTATCACAGAT	Internal protein-M23-peptidase	48_CRISPR_2_NODE_1554	26558	26594	26564; 26567; 26582	–	–	–	–	–
58	AGAAGTTTTTCTTCGCCAA GATATATGGTGCTGGTCTAA	DNA polymerase	58_CRISPR_10_NODE_13413	282	320	–	282; 292; 313	G/T; C/A; G/A	0.079; 0.552; 0.549	All silent	CGA/AGA; GGG/CGT; TTC/TTT
58	GTGTTGGTGCTCTTGGAGT ACCGTTCAGAATAGGT	Hypothetical protein	58_CRISPR_10_NODE_13413	35908	35942	–	35908	G/A	0.052	Silent	GGC/GGT
58	AGTTGTGCCCCTTGAGCTA GAGAATTTGCTGCACCT	Internal protein-M23-peptidase	58_CRISPR_10_NODE_13413	24692	24727	–	–	–	–	–	–
58	TAAACTGGTCGGGATTGTG TACATTCCATGCACTC	NC	58_CRISPR_10_NODE_13413	8740	8774	–	8753	C/G	0.53	–	–
58	ACTATCTGATCAAACCGGG GCTACACGGTAAATCGTTAGA	Tail fiber protein	58_CRISPR_10_NODE_13413	36649	36688	36650	36675	C/T	0.522	A/V	GCT/GTT
58	ACCTTTCAGACCTAACTCT AAAGTTACTATCACAGAT	Internal protein-M23-peptidase	58_CRISPR_5_NODE_1091	26558	26594	26565; 26567; 26582	–	–	–	–	–
58	CCCAACAACGTCTAAATAAA TCTTTCTATGATATGC	Hypothetical protein	58_CRISPR_8_NODE_4438	24219	24254	24226; 24229; 24238	–	–	–	–	–
58	AATACGGTTGTAGTACTCTT GAAGAGGTGTTACCG	Hypothetical protein	58_CRISPR_8_NODE_4438	30971	31005	30972	30978	G/A	0.588	Silent	ACG/ACA
58	GAAAGGGTAAGGTGTCAAA ATTGGGATTATTAGTGTTAG	Internal protein-M23-peptidase	58_CRISPR_8_NODE_4438	27172	27210	–	27180; 27181; 27186	T/A; C/A; A/C	0.626; 0.613; 0.575	S/T; S/Y; T/P	TCT/ACT; TCT/TAT; ACC/CCC
58	GCATTAATCGCGGGGTTAG GGTGATACCACCTA	Tail protein	58_CRISPR_8_NODE_4438	34211	34243	34214; 34226; 34241	–	–	–	–	–
58	TAGCTTAACATTACCACAG GGGATAAGCTGTTGTATATCC	Deoxycytidine triphosphate deaminase	58_CRISPR_9_NODE_4711	38225	38264	38264	38225; 38261	C/T G/A	0.057; 0.046	All silent	CTG/CTA; GAC; GAT
58	GACTTGATCTTTTCCGCT TTCTTGTAGCGCAGTATCTT	DNA polymerase	58_CRISPR_9_NODE_4711	668	705	–	671; 673	C/T; T/G	0.031; 0.040	R/K; Silent	AGG/AAG
58	ACGGGGTTGATCTTCCCCG CGAAGTGGTTGTCACCGAAT	dTMP kinase	58_CRISPR_9_NODE_4711	6338	6376	6375	6350	G/C	0.575	K/N	AAG/AAC

(Continued)

TABLE 2 | Continued

T°C Sample	Virotope Sequence	Viral target	CRISPR *loci*	Proto-spacer Start	Proto-spacer End	Mismatch position	SNV position	Alleles *	Frequency of in CRISPR allele	SNV effect	Codon change
58	AAATACATCCCCCACTTTAG GAGGTAACCCCAC	Hypothetical protein	58_CRISPR_9_NODE_4711	36127	36159	–	–	–	–	–	–
58	ACAGCGAAAGCAATTTGTC TCTGAGGCTAACAAGTT	Internal protein-M23-peptidase	58_CRISPR_9_NODE_4711	25742	25777	25776	–	–	–	–	–
58	GTCGTATCTCAATGTACTCT TTGTAGTCTTTCCA	Internal protein-M23-peptidase	58_CRISPR_9_NODE_4711	25917	25950	–	25946	C/A	0.041	Silent	ATC/ATA
58	CAATCACACCTAACCCCAT AGGTGACCGCACAACA	Portal protein	58_CRISPR_9_NODE_4711	15920	15954	–	15941; 15953	A/G; A/T	0.651; 0.055	All silent	GGA/GGG; ATA/ATT
58	TAGCTGATTGGAAAGCAGA CGCTGGATTATTACAC	Tail protein	58_CRISPR_9_NODE_4711	33859	33893	–	–	–	–	–	–
66	ATCTGTGATAGTAACTTTAG AGTTAGGTCTGAAAGGT	Internal protein-M23-peptidase	66_CRISPR_2_NODE_1045	26558	26594	26566; 26567; 26582	–	–	–	–	–
66	TTAGACCAGCACCATATATC TTGGCGAAGAAAAACTTCT	DNA polymerase	66_CRISPR_3_NODE_1491	282	320	282	292; 313	C/A; G/A	0.437; 0.024	All silent	GGG/CGT; TTC/TTT
66	ACCTATTCTGAACGGTACTCCA AGAGCACCAACAC	Hypothetical protein	66_CRISPR_3_NODE_1491	35908	35942	35908	–	–	–	–	–
66	AGGTGCAGCAAATTCTCTAGC TCAAGGGGCACAACT	Internal protein-M23-peptidase	66_CRISPR_3_NODE_1491	24692	24727	–	–	–	–	–	–
66	GAGTGCATGGAATGTACA CAATCCCGACCAGTTTA	NC	66_CRISPR_3_NODE_1491	8740	8774	–	8753	C/G	0.051	–	–
66	TCTAACGATTTACCGTGTAGC CCCGGTTTGATCAGATAGT	Tail fiber protein	66_CRISPR_3_NODE_1491	36649	36688	36650	36675	C/T	0.056	A/V	GCT/GTT

**Virus allele/CRISPR allele.*

within these communities (Heidelberg et al., 2009; Davison et al., 2016). Thus far, no study has characterized viral composition and activity, or the identity of any complete viral genome. Here, using metagenomic and metatranscriptomic approaches, the composition of the most abundant and active viruses associated with the dominant members of the thermophilic bacterial community have been characterized, describing for the first time a full genome from a thermophilic cyanopodovirus (TC-CHP58). Moreover, the active cross-fire between this new cyanophage and its host is demonstrated, through TC-CHP58 population diversification (SNV), and *Mastigocladus* spp. CRISPR heterogeneity, as a response to selective pressure from the host defense system and viral predation, respectively.

Active and Ubiquitous Cyanophage-Type Caudovirales in Phototrophic Microbial Mats

The taxonomic classification of small subunit rRNA (**Supplementary Table S1**) indicates that the phototrophic mats in Porcelana hot spring are dominated by Bacteria (96% on average) as commonly observed in other thermophilic phototrophic microbial mats (Inskeep et al., 2013; Bolhuis et al., 2014).

Porcelana microbial mats are mainly built by filamentous representatives of two phototrophic phyla, Cyanobacteria (oxygenic) and Chloroflexi (anoxygenic), with *Mastigocladus*, *Chloroflexus,* and *Roseiflexus* as the main genera, respectively. This is verified by previous surveys carried out by the authors (Mackenzie et al., 2013; Alcamán et al., 2015), as well as investigations from the White Creek, Mushroom, and Octopus hot springs in Yellowstone (Miller et al., 2009; Inskeep et al., 2013; Klatt et al., 2013; Bolhuis et al., 2014), presenting similar pH, thermal gradient and low sulfide concentrations.

Porcelana dominant viruses (∼70% and ∼68% of metagenomic and metatranscriptomic reads) are from the families Myoviridae, Podoviridae, and Siphoviridae within the Caudovirales Order (**Figure 2**), which typically infect Bacteria and some non-hyperthermophilic Archaea (Maniloff and Ackermann, 1998). These results were also supported by TEM images (**Figure 1**). The small decrease in transcripts associated to caudovirales with the increase in temperature is due to the reduction of sequences related to Podovirus and Myovirus families. A plausible explanation, is that at high temperatures some representatives of these families might have a lysogenic lifestyle, then a fraction of them will remain inactive as prophages.

Dominance by Caudovirales was only reported recently from the Brandvlei hot spring, South Africa, a slightly acidic (pH 5.7) hot spring with moderate temperature (60°C) and green microbial mat patches (Zablocki et al., 2017). Previously, the presence of this viral order had only been suggested in moderate thermophilic phototrophic mats from Yellowstone hot springs, through indirect genomic approximations, such as spacers in CRISPR *loci*, from dominant bacterial members (Heidelberg et al., 2009; Davison et al., 2016) or classifications based on nucleotide motives in metaviromic data (Pride and Schoenfeld, 2008; Davison et al., 2016).

Contributions from megavirus sequences were also identified in Porcelana hot spring (**Figure 2**), with an average of ∼24% viral metagenomic reads, associated with unicellular eukaryotic hosts such as those from Phycodnaviridae and Mimiviridae families, and also the family Marseilleviridae, but to a lesser extent. The presence of VLPs from these three viral families could not be corroborated through TEM, using the limited available viral fraction ($<0.2\ \mu m$) within the community, as it has been previously documented that nucleocytoplasmic large DNA viruses (NCLDV) particles are only found in larger viral fractions (Pesant et al., 2015). The ubiquity of NCLDVs in hot springs was previously described in a hydrothermal freshwater lake in Yellowstone, with assemblies of genomes from Phycodnaviridae and Mimiviridae (Zhang et al., 2015).

Viral relative abundances and activity reported here can be affected by the lack of replicates at this highly local heterogeneity samples. However, the fact of having three different temperature sampling points for metagenomics and metatranscriptomics, partially compensates the replicate limitation.

Furthermore, many viruses in an environmental sample share a degree of similarity in their genomic sequence, and this intrinsic complexity of metagenomic/metatranscriptomic samples makes difficult to accurately estimate the relative abundances or activity of specific phages at low ranks of taxonomy tree, such as the species level (Sohn et al., 2014). To avoid this problem, our strategy focused on the use of the LCA algorithm at higher taxonomic levels (Order and Family) to classify the viral reads, as well as for the inferred hosts, we use the phylum level.

Virus-host inference in Porcelana phototrophic mats (**Figure 3B**), demonstrated that the most frequent targets for viral infections were the most dominant and active components of the bacterial communities. Similarly, this is the case in other environments, such as in the human microbiome (Macklaim et al., 2013) and marine communities (Thingstad et al., 2014; Zeigler-Allen et al., 2017). In Porcelana, it is demonstrated that within microbial mats at 48°C and 58°C, cyanophages were among the most active viruses (**Figure 3B**), as were Cyanobacteria, such as *Mastigocladus* spp., as exemplified in terms of primary production and nitrogen fixation (Alcamán et al., 2015). The presence of cyanophages has been previously suggested in Yellowstone hot spring phototrophic mats (Heidelberg et al., 2009; Davison et al., 2016), and more recently in the Brandvlei hot spring, South Africa (Zablocki et al., 2017). Heidelberg et al. (2009) found that CRISPR spacers in unicellular cyanobacteria *Synechococcus* isolates (Syn OS-A and Syn OS-B9) from Octopus Hot Spring, might have 23 known viral targets (lysozyme-related reads, PFAM DUF847) on an independently published metavirome from the same hot spring. More recently, 171 viral contigs associated with the host genus S*ynechococcus*, based on tetranucleotide frequencies, were identified from a microbial mat (60°C) metavirome from Octopus Spring. The majority of the annotated ORFs on the viral contigs coded for glycoside hydrolases, with lysozyme activity, identifying six CRISPR proto-spacers in those genes (Davison et al., 2016). Even though a taxonomic

relationship with cyanophages was not confirmed for those proto-spacers containing contigs (Heidelberg et al., 2009; Davison et al., 2016), it provides evidence toward the presence of cyanophages related sequences within these thermophilic mats. The work by Zablocki et al. (2017) reconstructed a 10 kb partial genome of a new cyanophage (BHS3) from Brandvlei hot spring metavirome, stating that cyanophages appear to be the dominant viruses in the hot spring. The BHS3 contig (MF098555) contains nine ORFs, with the majority of the identified proteins having a close relation to the Cyanophage PP and *Phormidium* phage Pf-WMP3, which infect freshwater filamentous cyanobacteria *Phormidium* and *Plectonema*.

The presence of cyanophages related sequences in thermophilic phototrophic mats is significant, since these viruses are known to play an important role in the evolution of cyanobacteria (Shestakova and Karbysheva, 2015). Cyanophages affect the rate and direction of cyanobacterial evolutionary processes, through the regulation of abundance, population dynamics, and natural community structure. This has been extensively studied and demonstrated for marine environments (Weinbauer and Rassoulzadegan, 2004; Avrani et al., 2011). These cyanophages are proven to play a relevant role in the marine biogeochemical cycles, through the infection and lysis of Cyanobacteria, affecting carbon and nitrogen fixation (Suttle, 2000). Moreover, cyanophages act as a global reservoir of genetic information, as they are vectors for gene transfer, meaning that cyanobacteria can acquire novel attributes within aquatic environments (Kristensen et al., 2010; Chénard et al., 2016).

Caudoviruses were prevalent at 66°C in Porcelana, and potentially infecting Firmicutes, Proteobacteria, and Actinobacteria. These phila have also been previously identified in other hot springs at temperatures above 76°C, such as in Octopus and Bear Paw (Pride and Schoenfeld, 2008). At high temperatures in Porcelana also the phylum Chloroflexi was dominant in the phototrophic mat (**Figure 3A**). However, viral sequences related to this taxon could not be retrieved, as neither viruses nor viral sequences have been confirmed to infect members of this phylum in any environment. Davison et al. (2016), described viral contigs associated with *Roseiflexus* sp. from a metavirome from Octopus Spring, but only raw reads are publicly available, without taxonomic assignation. Finally, the recently released IMG/VR database (Paez-Espino et al., 2016) contains three contigs associated by CRISPR spacers to *Chloroflexus* sp. Here, a BLASTP analyses against RefSeq viral proteins revealed that six of these proteins have a best hit in *Mycobacterium* phage proteins and one which best hit was a *Clavibacter* phage protein. These findings, suggest that some of the viral reads classified as Actinobacteria viruses could be instead from unknown Chloroflexi viruses.

Viral Mining Reveals a New Infective Thermophilic Cyanopodovirus Lineage

Metagenomic surveys of viral genomes are an effective way to detect unknown viruses (Roux et al., 2015a,b; Zhang et al., 2015; Voorhies et al., 2016). In metagenomics, two key elements for virus detection are the presence of viral hallmark genes and the circularity of viral contigs (Roux et al., 2015a,b). Based on these two principles, a complete genome (TC-CHP58) was identified. The genome was represented by a viral contig of 50 kb, which is a typical size for Caudovirales members from the Podoviridae family. The genome size and viral core proteins affiliated with the Podovirus seems to make TC-CHP58 the first report of a full genome of a thermophilic cyanopodovirus. Moreover, the genome organization (**Figure 4B**) shows a consistent synteny with other cyanopodoviruses, which also lack RNA polymerase inside the T7 supergroup, as described for the viruses Pf-WMP4, Pf-WMP3, Cyanophage PP, Anabaena phage A-4L (Liu et al., 2007, 2008; Zhou et al., 2013; Ou et al., 2015), and the recently reported partial genome of the thermophilic BHS3 cyanophage (Zablocki et al., 2017). Initially, the presence of a single-subunit RNA polymerase that binds phage specific promoters was considered to be a major, and unique characteristic of the T7 supergroup (Dunn et al., 1983). However, more recently, it has been proposed that podoviruses that share extensive homology with T7, but lack the phage RNA polymerase, are still part of the T7 supergroup, as distant and probably ancient branches (Hardies et al., 2003).

TC-CHP58 presented a genome organization that can be divided into two portions (**Figure 4A**); with ORFs in the sense strand related to DNA replication and modification, and genes encoded in the antisense strand related to virion assembly. This genome organization is also present in other freshwater T7-related podoviruses that infect filamentous cyanobacteria (Liu et al., 2007, 2008; Zhou et al., 2013; Ou et al., 2015), including the thermophilic BHS3 cyanophage (Zablocki et al., 2017). This setup is also similar to the class II and III organization genes in T7-like viruses, where class II genes are responsible for DNA replication and metabolism, and class III genes include structural and maturation genes (Dunn et al., 1983). The VIRFAM analysis of neck protein organization verifies the classification of TC-CHP58 within the Podoviridae family (**Supplementary Figure S3**), where the Type 3 podovirus encompasses T7-like phages from Autographivirinae subfamilies and several other genera (Lopes et al., 2014). The T7-like classification for TC-CHP58, and other podoviruses that infect freshwater filamentous cyanobacteria, is supported by the organization of the genome into two portions as well as the organization of the neck proteins.

The phylogenetic position of TC-CHP58, based on DNA polymerase I (DNApol) (**Figure 6**) and MCP (**Supplementary Figure S4**) predicted proteins, confirm the affiliation of this new virus within the family Podoviridae. Both phylogenetic markers verify the separation between the marine from the freshwater cyanopodoviruses within the T7 family, as previously proposed (Liu et al., 2007; Ou et al., 2015). These results also support the connection between the T7 phages and marine and freshwater cyanopodoviruses (Chen and Lu, 2002; Hardies et al., 2003; Liu et al., 2007; Ou et al., 2015), including TC-CHP58 and BHS3 as representatives of a novel, and potentially globally distributed thermophilic cyanophage lineage. Moreover, this data demonstrates that marine and freshwater cyanopodoviruses, including the thermophilic TC-CHP58, are part of the Autographivirinae subfamily as previously suggested

for Cyanophage P60 and Roseophage SIO1 (Labonté et al., 2009), both included in this analysis.

In Porcelana, the virus host ratio relating to TC-CHP58 presence was lower than the typical values observed in freshwater environments (Maranger and Bird, 1995), being more similar to other geothermal environments where viral density is typically lower, with 10- to 100-fold less viruses than host cells (López-López et al., 2013). This is expected, considering that there are abundant cyanobacteria in phototrophic mats in Porcelana in comparison with the 10^4 mL^{-1} VLPs observed in the water of hot springs (Breitbart et al., 2004). It is also demonstrated that TC-CHP58 presented higher infection efficiency, as revealed by the viral DNA to RNA ratios at lower temperatures (58°C, then 48°C) with cyanobacteria dominating, while at 66°C most of the TC-CHP58 remained inactive (**Figure 5**). Infection inefficiency is multidimensional, as it initiates from reduced phage adsorption, RNA, DNA, and protein production (Howard-Varona et al., 2017). Thus, the high copy number of TC-CHP58 DNA at 66°C may be due to the persistence of viral DNA (Mengoni et al., 2005) encapsidated extracellularly and intermixed in the microbial mat were the host (*Mastigocladus* spp.) has a low activity as evidenced by the low expression of the RUBISCO gene and the CRISPR *loci*. An alternative explanation is the absence, or the diminished presence, of the specific host due to intraspecific diversification as evidenced by the existence of different CRISPR *loci* at different temperatures. This theory has been proposed for other cyanobacteria, such as *Prochlorococcus* and *Phormidium*, where slight differences in fitness, niche, and selective phage predation, explain the coexistence of different populations (Kashtan et al., 2014; Voorhies et al., 2016). The last explanation acquires special importance in light of recent evidence that variations in the structure and function of the heterocyst and differential CRISPR *loci* are fundamental to diversification of *Mastigocladus laminosus* (also known as *Fischerella thermalis*), a cosmopolitan thermophilic cyanobacterium, reinforcing the importance of viral predation (Sano et al., 2018).

CRISPR Spacers Assign *Mastigocladus* spp. as Putative Hosts for TC-CHP58

It was possible to verify *Mastigocladus* spp. as putative hosts for the new cyanopodovirus (TC-CHP58), via the analysis of CRISPR spacers found in the cyanobacteria, recovered from contigs obtained in the same metagenomic datasets. This methodology has been previously used for the identification of novel viruses in hot springs (Heidelberg et al., 2009; Snyder et al., 2010; Davison et al., 2016), as well as in other environments such as acid mines (Andersson and Banfield, 2008), the human microbiome (Stern et al., 2012), as well as sea ice and soils (Sanguino et al., 2015).

Observations from the CRISPR *loci* over all temperatures (**Table 2**) indicated that, in general, proto-spacers in the TC-CHP58 genome were distributed on coding, and therefore more conserved regions. The expression of seven CRISPR *loci* (**Figure 5**), demonstrated the activity of the *Mastigocladus* spp. defense system against TC-CHP58 over all temperatures. CRISPR arrays are transcribed into a long precursor, containing spacers and repeats, that are processed into small CRISPR RNAs (crRNAs) by dedicated CRISPR-associated (Cas) endoribonucleases (Brouns et al., 2008). Although it is not possible to measure mature crRNAs, as due to their small size they are likely to be filtered out in RNA-seq libraries, this approximation has been validated using large datasets (Ye and Zhang, 2016).

Despite variations in the number of CRISPR *loci* observed at each temperature, with 60% of the total CRISPR *loci* found in Mastigocladus contigs at 58°C, the abundance of reads agreed with the abundance of other genes required by these cyanobacteria, such as the RUBISCO gene (**Figure 5**). This further verified that the *loci* are from *Mastigocladus* populations. The different CRISPR *loci* found over the different temperatures in Porcelana (**Table 2**), also reinforces the notion that diversification of *Mastigocladus* is partly due to selective pressure exerted by the predation of viruses, such as TC-CHP58. This theory has been previously put forward for *Mastigocladus laminosus* in Yellowstone (Sano et al., 2018), and proposed for marine cyanobacteria (Rodriguez-Valera et al., 2009; Kashtan et al., 2014).

Furthermore, each CRISPR *loci* contains spacers that corresponds to different proto-spacers in the TC-CHP58 genome. Increases in spacer number and diversity against the same virus may explain the increase in interference, whilst decreasing the selection of escape mutants (Staals et al., 2016). Priming mechanisms are the most efficient form of obtaining new spacers (Staals et al., 2016), using a partial match between a pre-existing spacer and the genome of an invading phage to rapidly acquire a new "primed" spacer (Westra et al., 2016). Then, over-representation of spacer sequences in some regions of the TC-CHP58 genome may be related to a site that has already been sampled by the CRISPR-Cas machinery or by other biases such as the secondary structure of phage ssDNA, GC content, and transcriptional patterns (Paez-Espino et al., 2013).

The selection pressure of multiple spacers in *Mastigocladus* CRISPR *loci* leads to the emergence of SNVs in the TC-CHP58 viral populations (**Table 2**), which cause mismatches between spacers and proto-spacers, resulting in the attenuation or evasion of the host immune response (Shmakov et al., 2017). It is still possible to utilize mismatched spacers for interference and/or primed adaptation, however, the degree of tolerance to mismatches for interference among the CRISPR-Cas, varies substantially between different CRISPR-Cas type systems (Shmakov et al., 2017). The variable frequency (0.6–0.02) of the corresponding spacer SNVs alleles on TC-CHP58 proto-spacers, suggests that some variants are more prevalent throughout the population, regardless of whether the SNV causes a silent mutation. Based on this evidence, it has been proposed that, for other microbial communities, only the most recently acquired spacer can exactly match the virus. This suggests that community stability is driven by compensatory shifts in host resistance levels and virus population structure (Andersson and Banfield, 2008).

The present study describes the underlying viral community structure and activity of thermophilic phototrophic mats. Moreover, abundant virus populations are linked to dominant bacteria, demonstrating the effectiveness of omics approaches in

estimating the importance and activity of a viral community, in this case with thermophilic cyanophages.

Additionally, the first full genome of a new T7-related virus that infects thermophilic representatives of the cyanobacterium *Mastigocladus* spp. was here retrieved. This genome may represent a novel, globally present, freshwater thermophilic virus from a new lineage from the Podoviridae family. The latter was strongly suggested by the significant phylogenetic relationship and shared gene organization with the BHS3 cyanophage partial genome (South Africa). Even more, TC-CHP58 proteins also matches several contigs that include common viral hallmarks genes in the IMG/VR database. However, further work is necessary to fully understand the global representation and relevance of this virus, which complete genome is presented here as first reference available.

Finally, the evolutionary arms race between a specific cyanobacteria-cyanophage in the natural environment is exposed, where a there exist a variety of potential scenarios. For instance, host resistance may increase over time forcing the decrease of viral populations, or a specific virus population may occasionally become extremely virulent and cause the crash of the host population as proposed by the "kill the winner" model (Andersson and Banfield, 2008). Alternatively, if CRISPR systems and the diversification of the viral population remain in balance through time, a relatively stable virus and host community may result.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found NCBI as follow: Access to raw data for metagenomes and metatranscriptomes is available through NCBI BioProject ID PRJNA382437. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA382437. The genome of TC-CHP58 has the GenBank accession number KY888885. Contigs containing CRISPRs *loci* have been submitted to NCBI with GenBank accession numbers MG734911 to MG734917.

AUTHOR CONTRIBUTIONS

SG-L and BD conceived and designed the experiments. SG-L, OS, and FP performed the experiments. SG-L, CP-A, OS, FP, and BD analyzed the data. SG-L, CP-A, and BD wrote the paper.

FUNDING

This work was financially supported by Ph.D. scholarships CONICYT N° 21130667 and 21172022, and CONICYT grant FONDECYT N°1150171. Sequencing was funded by Spanish grant CTM2013-48292-C3-1-R.

ACKNOWLEDGMENTS

We are grateful to Huinay Scientific Field Station for making our work in the Porcelana hot spring possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2018.02039/full#supplementary-material

REFERENCES

Alcamán, M., Fernandez, C., Delgado, A., Bergman, B., and Díez, B. (2015). The cyanobacterium Mastigocladus fulfills the nitrogen demand of a terrestrial hot spring microbial mat. *ISME J.* 9, 2290–2303. doi: 10.1038/ismej.2015.63

Alcamán, M. E., Alcorta, J., Bergman, B., Vásquez, M., Polz, M., and 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. *Syst. Appl. Microbiol.* 40, 102–113. doi: 10.1016/j.syapm.2016.11.007

Andersson, A. F., and Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science (80-)* 320, 1047–1050. doi: 10.1126/science.1157358

Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature* 474, 604–608. doi: 10.1038/nature10172

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

Bhaya, D., Grossman, A. R., Steunou, A.-S., Khuri, N., Cohan, F. M., Hamamura, N., et al. (2007). Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.* 1, 703–713. doi: 10.1038/ismej.2007.46

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. *J. Virol.* 86, 5562–5573. doi: 10.1128/JVI.07196-11

Bolduc, B., Wirth, J. F., Mazurie, A., and Young, M. J. (2015). Viral assemblage composition in Yellowstone acidic hot springs assessed by network analysis. *ISME J.* 9, 1–16. doi: 10.1038/ismej.2015.28

Bolhuis, H., Cretoiu, M. S., and Stal, L. J. (2014). Molecular ecology of microbial mats. *FEMS Microbiol. Ecol.* 90, 335–350.

Breitbart, M., Wegley, L., Leeds, S., Rohwer, F., and Schoenfeld, T. (2004). Phage community dynamics in hot springs. *Appl. Environ. Microbiol.* 70, 1633–1640. doi: 10.1128/AEM.70.3.1633-1640.2004

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., et al. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science (80-)* 321:960 LP-964.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics* 10:1. doi: 10.1186/1471-2105-10-421

Chen, F., and Lu, J. (2002). Genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages genomic sequence and evolution of marine cyanophage P60: a new insight on lytic and lysogenic phages. *Appl. Environ. Microbiol.* 68, 2589–2594. doi: 10.1128/AEM.68.5.2589-2594.2002

Chénard, C., Wirth, J. F., and Suttle, C. A. (2016). Viruses infecting a freshwater filamentous cyanobacterium (*Nostoc* sp.) encode a functional CRISPR array and a proteobacterial DNA polymerase B. *mBio* 7:e00667–16. doi: 10.1128/mBio.00667-16

Cole, J. K., Peacock, J. P., Dodsworth, J. A., Williams, A. J., Thompson, D. B., Dong, H. L., et al. (2013). Sediment microbial communities in Great Boiling Spring

are controlled by temperature and distinct from water communities. *ISME J.* 7, 718–729. doi: 10.1038/ismej.2012.157

Crits-Christoph, A., Gelsinger, D. R., Ma, B., Wierzchos, J., Ravel, J., Davila, A., et al. (2016). Functional interactions of archaea, bacteria and viruses in a hypersaline endolithic community. *Environ. Microbiol.* 18, 2064–2077. doi: 10.1111/1462-2920.13259

Darriba, D., Taboada, G. L., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Davison, M., Treangen, T. J., Koren, S., Pop, M., and Bhaya, D. (2016). Diversity in a polymicrobial community revealed by analysis of viromes, endolysins and CRISPR spacers. *PLoS One* 11:e0160574. doi: 10.1371/journal.pone.0160574

Diemer, G. S., and Stedman, K. M. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA virus. *Biol. Direct.* 7, 1–14. doi: 10.1186/1745-6150-7-13

Dunn, J. J., Studier, F. W., and Gottesman, M. (1983). Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* 166, 477–535. doi: 10.1016/S0022-2836(83)80282-4

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* 35, 52–57. doi: 10.1093/nar/gkm360

Hardies, S. C., Comeau, A. M., Serwer, P., and Suttle, C. A. (2003). The complete sequence of marine bacteriophage VpV262 infecting *Vibrio parahaemolyticus* indicates that an ancestral component of a T7 viral supergroup is widespread in the marine environment. *Virology* 310, 359–371. doi: 10.1016/S0042-6822(03)00172-7

Heidelberg, J. F., Nelson, W. C., Schoenfeld, T., and 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:e4169. doi: 10.1371/journal.pone.0004169

Howard-Varona, C., Roux, S., Dore, H., Solonenko, N. E., Holmfeldt, K., Markillie, L. M., et al. (2017). Regulation of infection efficiency in a globally abundant marine *Bacteriodetes* virus. *ISME J.* 11, 284–295. doi: 10.1038/ismej.2016.81

Huang, H. W., Mullikin, J. C., and Hansen, N. F. (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* 16:235. doi: 10.1186/s12859-015-0624-y

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN Community edition – Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e4957. doi: 10.1371/journal.pcbi.1004957

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., and Rusch, D. B. (2013). The YNP metagenome project: environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. *Front. Microbiol.* 4:67. doi: 10.3389/fmicb.2013.00067

Jonker, C. Z., van Ginkel, C., and Olivier, J. (2013). Association between physical and geochemical characteristics of thermal springs and algal diversity in Limpopo Province, South Africa. *Water SA* 2013, 95–104. doi: 10.4314/wsa.v39i1.10

Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science (80-)* 344, 416–420. doi: 10.1126/science.1248575

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Klatt, C. G., Inskeep, W. P., Herrgard, M. J., Jay, Z. J., Rusch, D. B., Tringe, S. G., et al. (2013). Community structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. *Front. Microbiol.* 4:e106. doi: 10.3389/fmicb.2013.00106

Klatt, C. G., Wood, J. M., Rusch, D. B., Bateson, M. M., Hamamura, N., Heidelberg, J. F., et al. (2011). Community ecology of hot spring cyanobacterial mats: predominant populations and their functional potential. *ISME J.* 5, 1262–1278. doi: 10.1038/ismej.2011.73

Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *ISME J.* 11, 2399–2406. doi: 10.1038/ismej.2017.113

Koonin, E. V., Senkevich, T. G., and Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biol. Direct.* 1:27. doi: 10.1186/1745-6150-1-27

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., and Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* 18, 11–19. doi: 10.1016/j.tim.2009.11.003

Labonté, J. M., Reid, K. E., Suttle, C. A., Labont, J. M., Reid, K. E., and Suttle, C. A. (2009). Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. *Appl. Environ. Microbiol.* 75, 3634–3640. doi: 10.1128/AEM.02317-08

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

Lin, K.-H., Liao, B.-Y., Chang, H.-W., Huang, S.-W., Chang, T.-Y., Yang, C.-Y., et al. (2015). Metabolic characteristics of dominant microbes and key rare species from an acidic hot spring in Taiwan revealed by metagenomics. *BMC Genomics* 16:1029. doi: 10.1186/s12864-015-2230-9

Liu, X., Kong, S., Shi, M., Fu, L., Gao, Y., and An, C. (2008). Genomic analysis of freshwater cyanophage Pf-WMP3 infecting cyanobacterium *Phormidium foveolarum*: the conserved elements for a phage. *Microb. Ecol.* 56, 671–680. doi: 10.1007/s00248-008-9386-7

Liu, X., Shi, M., Kong, S., Gao, Y., and An, C. (2007). Cyanophage Pf-WMP4, a T7-like phage infecting the freshwater cyanobacterium *Phormidium foveolarum*: complete genome sequence and DNA translocation. *Virology* 366, 28–39. doi: 10.1016/j.virol.2007.04.019

Liu, Z., Klatt, C. G., Wood, J. M., Rusch, D. B., Ludwig, M., Wittekindt, N., et al. (2011). Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. *ISME J.* 5, 1279–1290. doi: 10.1038/ismej.2011.37

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* 16, 2659–2671. doi: 10.1111/1462-2920.12250

Lopes, A., Tavares, P., Petit, M. A., Guérois, R., and Zinn-Justin, S. (2014). Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* 15:1027. doi: 10.1186/1471-2164-15-1027.

López-López, O., Cerdán, M., and González-Siso, M. (2013). Hot spring metagenomics. *Life* 3, 308–320. doi: 10.3390/life3020308

Mackenzie, R., Pedrós-Alió, C., and Díez, B. (2013). Bacterial composition of microbial mats in hot springs in Northern Patagonia: variations with seasons and temperature. *Extremophiles* 17, 123–136. doi: 10.1007/s00792-012-0499-z

Macklaim, J. M., Fernandes, A. D., Di Bella, J. M., Hammond, J.-A., Reid, G., and Gloor, G. B. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* 1:12. doi: 10.1186/2049-2618-1-12

Maniloff, J., and Ackermann, H. W. (1998). Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. *Arch. Virol.* 143, 2051–2063. doi: 10.1007/s007050050442

Maranger, R., and Bird, D. F. (1995). Viral abundance in aquatic systems: a comparison between marine and fresh waters. *Mar. Ecol. Prog. Ser.* 121, 217–226. doi: 10.3354/meps121217

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10. doi: 10.14806/ej.17.1.200

Mengoni, A., Tatti, E., Decorosi, F., Viti, C., Bazzicalupo, M., and Giovannetti, L. (2005). Comparison of 16S rRNA and 16S rDNA T-RFLP approaches to study bacterial communities in soil microcosms treated with chromate as perturbing agent. *Microb. Ecol.* 50, 375–384. doi: 10.1007/s00248-004-0222-4

Miller, S. R., Purugganan, M., and Curtis, S. E. (2006). Molecular population genetics and phenotypic diversification of two populations of the thermophilic

cyanobacterium *Mastigocladus laminosus*. *Appl. Environ. Microbiol.* 72, 2793–2800. doi: 10.1128/AEM.72.4.2793-2800.2006

Miller, S. R., Strong, A. L., Jones, K. L., and 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. *Appl. Environ. Microbiol.* 75, 4565–4572. doi: 10.1128/AEM.02792-08

Ou, T., Liao, X. Y., Gao, X. C., Xu, X. D., and Zhang, Q. Y. (2015). Unraveling the genome structure of cyanobacterial podovirus A-4L with long direct terminal repeats. *Virus Res.* 203, 4–9. doi: 10.1016/j.virusres.2015.03.012

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

Paez-Espino, D., Morovic, W., Sun, C. L., Thomas, B. C., Ueda, K. I., Stahl, B., et al. (2013). Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat. Commun.* 4, 1430–1437. doi: 10.1038/ncomms2440

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from. *Cold Spring Harb. Lab. Press Method* 1, 1–31.

Pawlowski, A., Rissanen, I., Bamford, J. K. H., Krupovic, M., and Jalasvuori, M. (2014). Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. *Arch. Virol.* 159, 1541–1554. doi: 10.1007/s00705-013-1970-6

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* 2:150023. doi: 10.1038/sdata.2015.23

Prangishvili, D., and Garrett, R. A. (2004). Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem. Soc. Trans.* 32, 204–208. doi: 10.1042/bst0320204

Pride, D. T., and Schoenfeld, T. (2008). Genome signature analysis of thermal virus metagenomes reveals Archaea and thermophilic signatures. *BMC Genomics* 9:420. doi: 10.1186/1471-2164-9-420

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, 590–596. doi: 10.1093/nar/gks1219

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rachel, R., Bettstetter, M., Hedlund, B. P., Häring, M., Kessler, A., Stetter, K. O., et al. (2002). Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Arch. Virol.* 147, 2419–2429. doi: 10.1007/s00705-002-0895-2

Redder, P., Peng, X., Brügger, K., Shah, S. A., Roesch. F., Greve, B., et al. (2009). Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. *Environ. Microbiol.* 11, 2849–2862. doi: 10.1111/j.1462-2920.2009.02009.x

Richter, M., Rosselló-Móra, R., Glöckner, F., and Peplies, J. (2016). JSpeciesWS: a web server for prokaryotic species circumscription based on pairwise genome comparison. *Bioinformatics* 32, 929–931. doi: 10.1093/bioinformatics/btv681

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasiæ, L., Thingstad, T. F., Rohwer, F., et al. (2009). Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* 7, 828–836. doi: 10.1038/nrmicro2235

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. doi: 10.1093/sysbio/sys029

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

Roux, S., Hallam, S. J., Woyke, T., and Sullivan, M. B. (2015b). Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4:e08490. doi: 10.7554/eLife.08490

Sanguino, L., Franqueville, L., Vogel, T. M., and Larose, C. (2015). Linking environmental prokaryotic viruses and their host through CRISPRs. *FEMS Microbiol. Ecol.* 91, 1–9. doi: 10.1093/femsec/fiv046

Sano, E. B., Wall, C. A., Hutchins, P. R., and Miller, S. R. (2018). Ancient balancing selection on heterocyst function in a cosmopolitan cyanobacterium. *Nat. Ecol. Evol.* 2, 510–519. doi: 10.1038/s41559-017-0435-9

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schmieder, R., Lim, Y. W., and Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics* 28, 433–435. doi: 10.1093/bioinformatics/btr669

Schoenfeld, T., Patterson, M., Richardson, P. M., Wommack, K. E., Young, M., and Mead, D. (2008). Assembly of viral metagenomes from yellowstone hot springs. *Appl. Environ. Microbiol.* 74, 4164–4174. doi: 10.1128/AEM.02598-07

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shestakova, S. V., and Karbysheva, E. A. (2015). The role of viruses in the evolution of Cyanobacteria. *Biol. Bull. Rev.* 5, 527–537. doi: 10.1134/S2079086415060079

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., and Koonin, E. (2017). The CRISPR spacer space is dominated by crossm the CRISPR spacer space is dominated. *mBio* 8, 1–18. doi: 10.1128/mBio.01397-17

Snyder, J. C., Bateson, M. M., Lavin, M., and Young, M. J. (2010). Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. *Appl. Environ. Microbiol.* 76, 7251–7258. doi: 10.1128/AEM.01109-10

Sohn, M. B., An, L., Pookhao, N., and Li, Q. (2014). Accurate genome relative abundance estimation for closely related species in a metagenomic sample. *BMC Bioinformatics* 15:242. doi: 10.1186/1471-2105-15-242

Staals, R. H. J., Jackson, S. A., Biswas, A., Brouns, S. J. J., Brown, C. M., and Fineran, P. C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat. Commun.* 7:12853. doi: 10.1038/ncomms12853

Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* 22, 1985–1994. doi: 10.1101/gr.138297.112

Steunou, A.-S., Bhaya, D., Bateson, M. M., Melendrez, M. C., Ward, D. M., Brecht, E., et al. (2006). In *situ* analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. *Proc. Natl. Acad. Sci. U.S.A.* 103, 2398–2403. doi: 10.1073/pnas.0507513103

Steunou, A.-S., Jensen, S. I., Brecht, E., Becraft, E. D., Bateson, M. M., Kilian, O., et al. (2008). Regulation of *nif* gene expression and the energetics of N_2 fixation over the diel cycle in a hot spring microbial mat. *ISME J.* 2, 364–378. doi: 10.1038/ismej.2007.117

Stewart, W. (1970). Nitrogen fixation by blue-green algae in Yellowstone thermal areas. *Phycologia* 9, 261–268. doi: 10.2216/i0031-8884-9-3-261.1

Suttle, C. A. (2000) "Cyanophages and their role in the ecology of cyanobacteria," in *The Ecology of Cyanobacteria*, eds B. A. Whitton and M. Potts (Dordrecht: Springer). doi: 10.1007/0-306-46855-7

Tekere, M., Lötter, A., Olivier, J., Jonker, N., and Venter, S. (2011). Metagenomic analysis of bacterial diversity of Siloam hot water spring, Limpopo, South Africa. *Afr. J. Biotechnol.* 10, 18005–18012.

Thingstad, T. F., Vage, S., Storesund, J. E., Sandaa, R.-A., and Giske, J. (2014). A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc. Natl. Acad. Sci. U.S.A.* 111, 7813–7818. doi: 10.1073/pnas.1400909111

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., and Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* 44, 1–4. doi: 10.1093/nar/gkw256

Uldahl, K., and Peng, X. (2013). "Biology, biodiversity and application of thermophilic viruses," in *Thermophilic Microbes in Environmental and Industrial Biotechnology*, eds T. Satyanarayana and K. Y. Littlechild Jennifer (Berlin: Springer), 271–306. doi: 10.1007/978-94-007-5899-5_10

Van der Meer, M. T., Klatt, C. G., Wood, J., Bryant, D. A., Bateson, M. A., Lammerts, L., et al. (2010). Cultivation and genomic, nutritional, and lipid biomarker characterization of roseiflexus strains closely related to predominant in situ populations inhabiting yellowstone hot

spring microbial mats. *J. Bacteriol.* 12, 3033–3042. doi: 10.1128/JB.01610-09

Voorhies, A. A., Eisenlord, S. D., Marcus, D. N., Duhaime, M. B., Biddanda, B. A., Cavalcoli, J. D., et al. (2016). Ecological and genetic interactions between cyanobacteria and viruses in a low-oxygen mat community inferred through metagenomics and metatranscriptomics. *Environ. Microbiol.* 18, 358–371. doi: 10.1111/1462-2920.12756

Weinbauer, M. G., and Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* 6, 1–11. doi: 10.1046/j.1462-2920.2003.00539.x

Westra, E. R., Dowling, A. J., Broniewski, J. M., and van Houte, S. (2016). Evolution and Ecology of CRISPR. *Ann. Rev. Ecol. Evol. Syst.* 47, 307–331. doi: 10.1146/annurev-ecolsys-121415-032428

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201. doi: 10.1093/nar/gks918

Wu, Y. W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605–607. doi: 10.1093/bioinformatics/btv638

Ye, Y., and Zhang, Q. (2016). Characterization of CRISPR RNA transcription by exploiting stranded metatranscriptomic data. *RNA* 22, 945–956. doi: 10.1261/rna.055988.116

Yu, M. X., Slater, M. R., and Ackermann, H.-W. (2006). Isolation and characterization of *Thermus* bacteriophages. *Arch. Virol.* 151, 663–679. doi: 10.1007/s00705-005-0667-x

Zablocki, O., van Zyl, L. J., Kirby, B., and Trindade, M. (2017). Diversity of dsDNA viruses in a South African hot spring assessed by metagenomics and microscopy. *Viruses* 9:348. doi: 10.3390/v9110348

Zeigler-Allen, L., McCrow, J. P., Ininbergs, K., Dupont, C. L., Badger, J. H., Hoffman, J. M., et al. (2017). The baltic sea virome: diversity and transcriptional activity of DNA and RNA viruses. *mSystems* 2, e125-16. doi: 10.1128/mSystems.00125-16

Zhang, W., Zhou, J., and Wang, Y. (2015). Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. *Sci. Rep.* 5:15131. doi: 10.1038/srep15131

Zhou, Y., Lin, J., Li, N., Hu, Z., and Deng, F. (2013). Characterization and genomic analysis of a plaque purified strain of cyanophage PP. *Virol. Sin.* 28, 272–279. doi: 10.1007/s12250-013-3363-0

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