UNIVERSITAT DE BARCELONA

Interactions between marine picoeukaryotes and their viruses one cell at a time

Interacciones entre picoeucariotas marinos y sus virus célula a célula

Yaiza M. Castillo de la Peña

Interactions between marine picoeukaryotes and their viruses one cell at a time

(Interacciones entre picoeucariotas marinos y sus virus célula a célula)

Yaiza M. Castillo de la Peña

Tesis doctoral presentada por Dª Yaiza M. Castillo de la Peña para obtener el grado de Doctora por la Universitat de Barcelona y el Institut de Ciències del Mar, programa de doctorado en Biotecnología, Facultat de Farmàcia i Ciències de l'Alimentació.

Directoras: Dra. Mª Dolors Vaqué Vidal y Dra. Marta Sebastián Caumel

Tutora: Dra. Josefa Badía Palacín

Universitat de Barcelona (UB)

Institut de Ciències del Mar (ICM-CSIC)

La doctoranda	La directora	La co-directora	La tutora
Yaiza M. Castillo	Dolors Vaqué	Marta Sebastián	Josefa Badía

En Barcelona, a 25 de noviembre de 2019

Cover design and images: © Yaiza M. Castillo.
TEM images from Derelle *et al*., 2008

This thesis has been funded by the Spanish Ministry of Economy and competitivity (MINECO) through a PhD fellowship to Yaiza M. Castillo de la Peña (BES-2014-067849), under the program "Formación de Personal Investigador (FPI)", and adscribed to the project: "Impact of viruses on marine microbial communities using virus-host models and metagenomic analyzes." MEFISTO (Ref. CTM2013-43767-P, P.I.Dr. Dolors Vaqué).

Other projects that contributed partially to the completion of this thesis were: ALLFLAGS (CTM2016-75083-R, MINECO, P.I. Dr. Ramon Massana), INTERACTOMICS (CTM2015-69936-P, MINECO, Dr. Ramiro Logares), the EU project SINGEK (H2020- MSCAITN-2015-675752, P.I. Dr. Ramon Massana) and *Tara* Oceans Expedition (http://www.embl.de/tara-ocean s/).

A mis padres y a mis abuelos

ACKNOWLEDGEMENTS

Primero de todo quiero agradecer a mis directoras de tesis Dolors Vaqué y Marta Sebastián por apoyarme y guiarme a lo largo de estos años de doctorado. Por su paciencia, motivación y conocimientos. Sin ellas esta tesis no habría sido posible.

Tampoco habría sido posible sin la inmensa ayuda de Irene Forn y Jeff Mangot, los cuales han trabajado conmigo codo con codo para poder sacar adelante este trabajo.

A todas mis compañeras del ICM que me han acompañado durante estos años y han hecho que la experiencia de la tesis tuviera un toque de luz especial. Gracias a Mariri Vicioso, Marta Masdeu, Idaira Santos, Carolina Marín, Elisabet Sà, Isabel Sanz, Anna Arias, Sergio González, Caterina Rodríguez, Dorleta Orue-Echevarría, Mireia Mestre, Ramiro Logares, Xavi Leal, Estela Romero, Isabel Marín, Néstor Arandia, Sheree Yau, Fran Aparicio y Sdena Nunes. Espero no haberme dejado a nadie, si es así, espero que me disculpéis.

Por último, gracias a mi familia por su apoyo incondicional. A mis padres, por todo el amor que me han dado y todo lo que han hecho por mí. A mis abuelos, por haber creído siempre en mí y haber estado tan orgullosos de su nieta, siempre estaréis en mi memoria. Os quiero.

CONTENTS

ABSTRACT

Marine viruses are key components of marine microbial communities, as they influence the cellular abundances and the community structure of microbes, participate in their genetic exchange, and intervene in the ocean biogeochemical cycles. Most studies dealing with the role of viruses in the marine environment have been done from a bulk community point of view, but going from the bulk community perspective to specific virus–host relationships is essential in order to understand the role of viruses in shaping a determined host community, in modifying host genomes, and ultimately in the release of organic compounds from the lysed cells. For this reason, in this thesis we implemented and applied different methodologies that are able to detect, visualize and quantify virus–host interactions in marine eukaryotes at the single cell level. We focused on picoeukaryotes (cells <3 µm) because they play crucial roles in marine food webs and biogeochemical cycles, and virus–host interactions in natural populations of these minute eukaryotes are largely unknown.

In the first chapter we combined previously developed techniques, used to assess prokaryotic host–phage interactions, to implement VirusFISH for detecting specific virus–host dynamics, using as a model system the photosynthetic picoeukaryote *Ostreoccocus tauri* and its virus OtV5. With the VirusFISH technique, we could also monitor the infection, as well as quantify the free viruses produced during the lysis of the host in a non-axenic culture, which allowed the calculation of the burst size. This study set the ground for the application of the VirusFISH technique to natural samples.

In the second chapter of this thesis, we applied VirusFISH to seawater samples from the Bay of Biscay (Cantabrian Sea) to study the dynamics of viral infection in natural populations of *Ostreococcus* along a seasonal cycle. We were able to quantify the percentage of cells infected over time, and compared these results with the transcriptional viral and host activities derived from metatranscriptomic

data. This constitutes the first study where a specific viral–host interaction has been visualized and monitored over time in a natural system.

Picoeukaryotes in the ocean are prevalently uncultured, and thus, in the third chapter of this thesis we went an step further to unveil novel viral–host relationships in eukaryotic uncultured hosts. For this purpose, we mined single amplified genomes (SAGs) of picoeukaryotes obtained during the *Tara* Oceans expedition for viral signatures. We found that almost 60% of the cells analyzed presented an associated virus with narrow host specificity. Some of the viral sequences were widely distributed and some geographically constrained, and they were preferentially found at the deep chlorophyll maximum. Moreover, we found a mavirus virophage potentially integrated in four SAGs of two different lineages, suggesting the presence of virophages is more common than previously thought.

In summary, in this thesis we have implemented and used techniques that allow us to detect and monitor specific virus–host interactions, which is one of the major challenges in marine viral ecology. On the one hand, VirusFISH arises as a powerful technique that can be easily adapted to any host–virus system that has been genome-sequenced. On the other hand, the results obtained with the single cell genomics offer the opportunity to formulate hypothesis based on detected viral–host interactions in uncultured prevalent marine picoeukaryotes, which can be later tested using experimental approaches.

RESUMEN

Los virus marinos son componentes clave de las comunidades microbianas ya que influencian las abundancias celulares y la estructura de las comunidades microbianas, participando en el intercambio genético e interviniendo en los ciclos biogeoquímicos en el océano. Se han realizado muchos estudios sobre el rol de los virus en ambientes marinos desde el punto de vista de la comunidad global. No obstante, es esencial para poder entender el rol que tienen los virus de "moldear" determinadas comunidades microbianas, modificar los genomas celulares y en última instancia liberar componentes orgánicos de las células lisadas al medio, que vayamos desde una visión más global de comunidad a una más específica de relación virus–hospedador. Por estas razones, en esta tesis implementamos y aplicamos diferentes metodologías para detectar, visualizar y cuantificar interacciones virus–hospedador a nivel de célula individual en eucariotas marinos. En este trabajo nos centramos en picoeucariotas (células de <3 µm) ya que se conoce muy poco de ellos a nivel de interacciones virus–hospedador en poblaciones naturales, a pesar de que juegan un papel crucial en las redes tróficas microbianas y en los ciclos biogeoquímicos.

En el primer capítulo combinamos técnicas desarrolladas previamente para la evaluación de interacciones procarióticas bacteriofago–hospedador, para implementar la técnica VirusFISH, que nos permite detectar dinámicas específicas virus–hospedador en poblaciones de eucariotas. Para ello usamos como modelo el sistema picoeucariótico fotosintético *Ostreococcus tauri* y su virus OtV5. Con la técnica de VirusFISH pudimos monitorizar la infección, así como cuantificar los virus libres producidos durante la lisis de los hospedadores en un cultivo no axénico, lo que nos permitió además calcular el tamaño de explosión (la cantidad de virus liberados por cada célula lisada). Este estudio estableció la base para la aplicación de VirusFISH en muestras naturales.

En el segundo capítulo de esta tesis, aplicamos VirusFISH en muestras de agua natural de la bahía de Vizcaya (Mar Cantábrico) para estudiar las dinámicas de

infección vírica en poblaciones naturales de *Ostreococcus*. Fuimos capaces de cuantificar el porcentaje de células infectadas durante un ciclo estacional y comparamos estos resultados con las actividades transcripcionales de virus y hospedadores derivadas de datos de metatranscriptómica. Este constituye el primer estudio donde se visualiza y monitoriza una interacción específica virus–hospedador a lo largo del tiempo en un sistema natural.

La mayor parte de los picoeucariotas en el océano no se pueden cultivar, por tanto, en el tercer capítulo de esta tesis nuestro objetivo fue descubrir nuevas relaciones virus–hospedador en células eucarióticas no cultivadas. Para este fin, analizamos genomas amplificados individuales (SAGs) de picoeucariotas obtenidos durante la campaña *Tara* Oceans. Encontramos que casi el 60% de las células analizadas presentaron al menos un virus, con una alta especificidad por el hospedador. Estas secuencias víricas se encontraron preferentemente en el máximo profundo de clorofila, estando algunas de ellas ampliamente distribuidas por los océanos y otras constreñidas geográficamente. Además, encontramos un virofago mavirus potencialmente integrado en cuatro SAGs de dos linajes distintos, sugiriendo que los virofagos son más comunes de lo que se pensaba anteriormente.

En resumen, en esta tesis hemos implementado y usado técnicas que nos han permitido detectar y monitorizar interacciones específicas virus–hospedador, lo cual es uno de los mayores retos en la ecología microbiana marina. Por un lado, VirusFISH surge como una técnica potente que puede ser fácilmente adaptada a cualquier sistema virus–hospedador del cual se tenga el genoma secuenciado. Por otro lado, los resultados obtenidos con la genómica de célula individual muestran la oportunidad de formular hipótesis basadas en interacciones virus–hospedador detectadas en picoeucariotas marinos no cultivados, que pueden ser posteriormente testadas mediante aproximaciones experimentales.

RESUM

Els virus marins són components clau de les comunitats microbianes ja que influencien les abundàncies cel·lulars i l'estructura de les comunitats microbianes participant en l'intercanvi genètic i intervenint en els cicles biogeoquímics en l'oceà. S'han realitzat molts estudis sobre el rol dels virus en ambients marins des d'un punt de vista de comunitat global. No obstant, és essencial per poder entendre el rol que tenen els virus de donar forma a determinades comunitats microbianes, modificar genomes cel·lulars i en última instància alliberar components orgànics de les cèl·lules lisades al medi, que anem des d'una visió més global de comunitat a una més específica de relació virus–hoste. Per aquestes raons, en aquesta tesi implementem i apliquem diferents metodologies per detectar, visualitzar i quantificar interaccions virus–hoste en eucariotes marins a nivell de cèl·lula individual. En aquest treball ens centrem en picoeucariotes (cèl·lules de <3 µm) ja que es coneix molt poc d'ells a nivell d'interaccions virus–hoste en poblacions naturals, tot i que juguen un paper crucial en les xarxes tròfiques microbianes i en els cicles biogeoquímics.

En el primer capítol combinem tècniques desenvolupades prèviament per la avaluació d'interaccions procariòtiques bacteriòfag–hoste, per implementar la tècnica VirusFISH, que ens permet detectar dinàmiques específiques de virus–hoste eucariòtics. Per això, fem servir com a model el sistema picoeucariòtic fotosintètic *Ostreococcus tauri* i el seu virus OtV5. Amb la tècnica de VirusFISH vam podem monitoritzar la infecció així com quantificar els virus lliures produïts durant la lisi dels hostes en un cultiu no axènic, lo qual ens va permetre a més calcular la grandària d'explosió (la quantitat de virus alliberats per cada cèl·lula lisada). Aquest estudi va establir la base per l'aplicació del VirusFISH en mostres naturals.

En el segon capítol d'aquesta tesi, apliquem el VirusFISH en mostres d'aigua natural de la badia de Vizcaya (Mar Cantàbric) per estudiar les dinàmiques de la

infecció vírica en poblacions natural d'*Ostreococcus* al llarg d'un cicle estacional. Vam ser capaços de quantificar el percentatge de cèl·lules infectades en el temps i vam comparar aquests resultats amb les activitats transcripcionals de virus i hostes derivades de dades de metatranscriptòmica. Aquest constitueix el primer estudi on es visualitza i monitoritza una interacció específica virus–hoste durant el temps en un sistema natural.

Els picoeucariotes en l'oceà són predominantment no cultivats, per tant, en el tercer capítol d'aquesta tesi vam anar un pas més enllà per descobrir noves relacions virus–hoste en hostes eucariòtics no cultivats. Vam analitzar genomes amplificats individualment (SAGs) de picoeucariotes obtinguts durant la campanya *Tara* Oceans per trobar senyals víriques que ens descobrissin noves associacions virus–host. Vam trobar que casi el 60% de les cèl·lules analitzades presentaven al menys un virus associat amb una estreta especificitat per l'hoste. Aquestes seqüencies víriques es van detectar preferentment al DCM, amb algunes d'elles distribuïdes àmpliament pels oceans i altres més limitades geogràficament. A més, vam trobar un virofag mavirus potencialment integrat en quatre SAGs en els que no es coneixia que existís aquesta relació.

En resum, en aquesta tesi hem implementat i utilitzat tècniques que ens han permès detectar i monitoritzar interaccions específiques virus–hoste, un dels majors reptes de l'ecologia marina microbiana. Per una banda, el VirusFISH sorgeix com una tècnica potent que pot ser fàcilment adaptada a qualsevol sistema virus–hoste del qual es tingui el genoma seqüenciat. Per altra banda, els resultats obtinguts amb la genòmica de cèl·lula individual mostren l'oportunitat de formular hipòtesis basades en interaccions virus–hoste detectades en picoeucariotes marins no cultivats, que poden ser posteriorment testats usant aproximacions experimentals.

GENERAL INTRODUCTION AND AIM OF THE THESIS

A GLOBAL INTRODUCTION TO VIRUSES

What is a virus?

Viruses (from the Latin word “virus”, meaning “poison”) are small infective particles composed by genetic material (single or double stranded DNA or RNA) protected by a protein coat called capsid, which sometimes is covered by a lipid envelope (Abedon, 2008).

The size of viruses generally ranges between 20 and 200nm and their observation requires epifluorescence or transmission electronic microscopes (Fig. 1). However, some recently discovered “giant viruses” can measure up to 750nm (e.g. Mimivirus, the biggest one discovered until now), and can be seen under a light microscope (Schrad *et al.*, 2017).

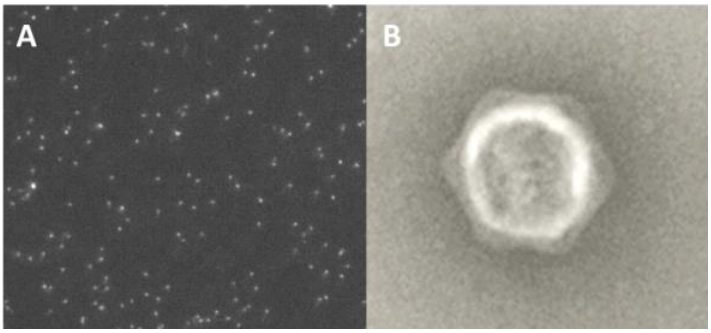


Figure 1. Micrographs of eukaryotic viruses using **A** epifluorescence microscopy (viruses stained with SYBRGold) and **B** transmission electronic microscopy. The higher resolution of the later allows the observation of the icosahedral shape of this particular virus. Micrographs acquired by YM Castillo at the Parc Científic de Barcelona.

Viral morphologies are varied: icosahedral, filamentous or head-tail (Fig. 2). Normally, viruses that infect eukaryotic cells are icosahedral or filamentous, while viruses that infect bacteria (called bacteriophages or phages) are head-tail (being the head the capsid, and the tail several proteins that some phages use to attach themselves to the cell).

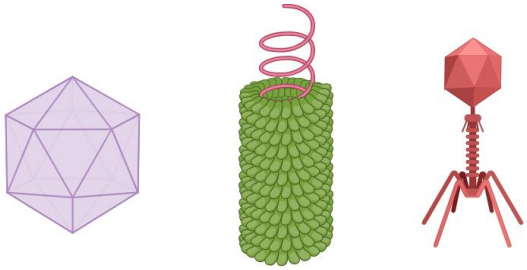


Figure 2. Representation of the most common morphologies known for viruses. Left: icosahedral capsid; middle: filamentous virus; right: head-tail phage (figure done with BioRender).

The viral genetic material encodes core (e.g. capsid proteins) and replication genes (e.g. polymerase), among other possibilities, but they do not present any genes related with their own metabolism (Maynard *et al.*, 2010). Therefore, viruses are unable to reproduce by themselves. To obtain a viral progeny, viruses use the machinery of a cell (called host) in their benefit. They stop the cell metabolism and replication activating those genes and machinery that transcribe the viral genes and assemble the viral structure (Goodwin *et al.*, 2015). Their range of action encompasses all type of cells i.e. prokaryotes (bacteria and archaea) and eukaryotes (from the smallest unicellular microorganisms as e.g. pico/nanoeukaryotes to pluricellular organisms as humans). However, despite there are millions of different types of viruses, nowadays only a little fraction is known, and from these, just a few genomes are well characterized (Bzhalava *et al.*, 2018).

Moreover, viruses are found in all Earth ecosystems where there is a cell and are the most abundant biological entities in the globe (~10^{31} viruses). If viruses were stretched end to end they would span ~10 million light years (Suttle, 2005). Thus, their role and high abundances make viruses a very important and critical component of our planet.

How viruses infect their hosts

In all cases, the mechanism of viral infection starts with the virus approaching and attaching to the host cell, where they recognize a cellular receptor. Later, depending on the virus behavior, they enter via endocytosis (typically eukaryotic viruses) (Yamauchi and Helenius, 2013; York, 2017) or inject the nucleic acid material into the cell (typically bacteriophages) (Grayson and Molineux, 2007; York, 2017).

After these common steps, depending on the type of virus or situation, there are several types of viral life cycles: lytic, lysogenic, pseudolysogenic and chronic, being the most common the lytic and lysogenic cycles (Abedon, 2008).

a. The lytic cycle

Lytic or virulent viruses lead to the host cell death through lysis (Fig. 3). During this cycle, the virus infects the cell and takes the metabolic machinery of the host on its behalf, producing new viruses (Echols, 1972). When the viral progeny is created, viruses produce lysine compounds that destroy the membrane or cell wall of prokaryotes (Pimentel, 2014) or eukaryotes (Daniels *et al.*, 2007), and viruses are released to the milieu bursting out the cell. Lytic viruses differ on the speed of assembly and release, but all converge in a fatal bursting of the cell (Echols, 1972).

b. The lysogenic cycle

Lysogenic or temperate viruses insert their nucleic acid into the host genome, becoming part of it, and remain as a silent virus in the cell. This virus is called “prophage” in the case of bacteria and “provirus” in the case of eukaryotes (Saussereau and Debarbieux, 2012; Filée, 2018). The prophage/provirus is transferred to the host progeny as its genome replicates together with the host genome, creating more genomic copies of the virus. When an environmental,

chemical or physical factor stresses the host, the virus is induced to revert the cycle from lysogenic to lytic (Fig. 3) (Echols, 1972).

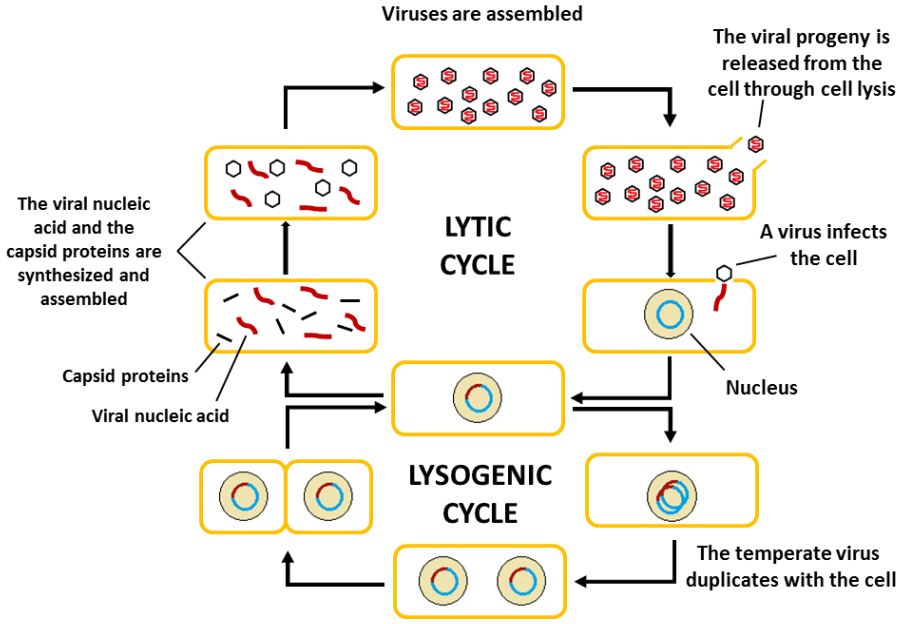


Figure 3. A simplified scheme of the lytic and lysogenic cycles. A virus can enter the lytic cycle and produce progeny or can be in lysogeny for many cellular duplications until it reverts to the lytic cycle (Drawing by YM Castillo).

c. The Pseudolysogenic cycle

This viral life strategy resembles the lysogenic cycle as in both cases the virus inserts its genome into the host and remains silent for several cellular generations. The difference between them lies in the place where the genome of the virus remains. While in the lysogenic cycle the viral genome is inserted into the host genome, in the pseudolysogenic cycle the viral genome remains in the cytoplasm (probably during a few cell generations) before it gets activated and produces the cell lysis (Fuhrman, 1999).

d. Chronic viruses

Chronic viruses follow the same strategy as lytic viruses with the exception that the progeny of viruses released by the host cell is non-lethal. Viruses are released by extrusion or budding, therefore, they do not kill the host (Fuhrman, 1999; Marciano, 1999). These types of viruses are the ones that can present a lipid envelope covering the capsid as they take part of the cellular membrane when they are released from the cell (Aloia *et al.*, 1993).

MARINE VIRUSES

General concepts

Marine viruses are the smallest and most abundant biological entities in the oceans, ranging from 10^4 to 10^7 viruses mL^{-1} of sea water (Suttle, 2005; Danovaro *et al.*, 2011). Their abundances represent 10 times the abundance of bacteria and approximatively 1000 times the planktonic protist abundance (Pernice *et al.*, 2015), and are positively correlated with biomass and activity of both bacteria and protist. Additionally, viral abundances decrease with the distance from shore and as we go down in the water column (Cochlan *et al.*, 1993).

Marine viruses are largely responsible for cell mortalities (bacterial, archaeal and protistan) in marine microbial communities (Munn, 2006) (Fig. 4), leading every day to ~10^{29} infection events (Brussaard *et al.*, 2008) and causing, on a daily basis, the lysis of ~20-40% of the prokaryotes and ~3% of the phytoplankton biomass standing stock in the oceans (Suttle, 1994, 2005). The lysis of phytoplankton cells causes that a fraction of the cellular carbon returns to the environment, avoiding its

transfer to higher trophic levels, in the process called viral shunt (Fuhrman, 1999) (Fig. 4).

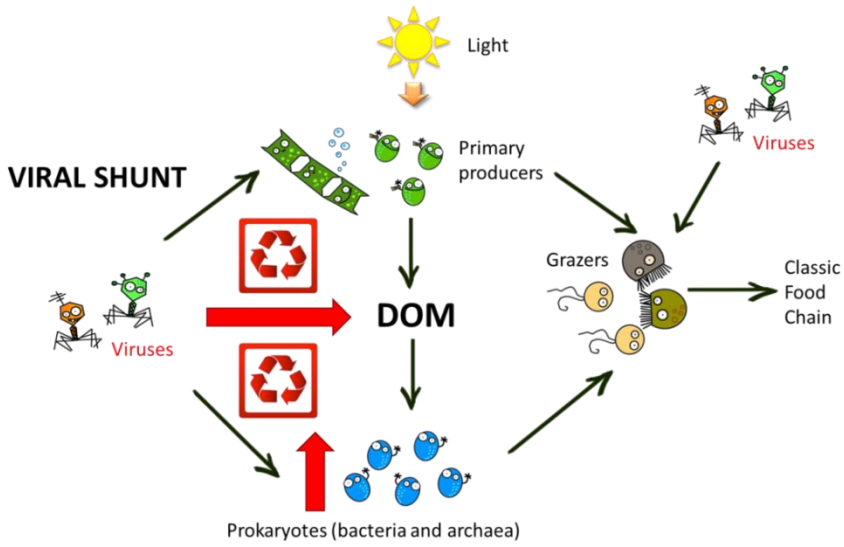


Figure 4. Microbial food web scheme in which the viral shunt is shown. Primary producers release dissolved organic matter (DOM) to the milieu. Heterotrophic prokaryotes use this DOM to grow. Both primary producers and heterotrophic prokaryotes are grazed by flagellates and ciliates, which in turn are grazed by higher microorganisms (the classic food chain). Viruses are involved in lysing the grazers, the heterotrophic prokaryotes and the primary producers. The lysis of the primary producers releases DOM to the milieu, which is used by the heterotrophic prokaryotes to grow, preventing the transfer of the phytoplankton carbon to higher trophic levels (to the classic food chain), and producing the recirculation and recycling of the biogenic carbon. This process is called **the viral shunt** (red arrows). Drawings done by Clara Ruíz-González.

By lysing their hosts, it has been estimated that viruses release to the marine environment 10^8-10^9 tons of biogenic carbon every day in the form of dissolved organic matter (DOM) (Suttle, 2005; Brussaard *et al.*, 2008; Lara *et al.*, 2017). Thus, viruses play important roles in marine biogeochemical cycles (Jover *et al.*, 2014), and also in population dynamics, because they control host abundances

and shape communities (Breitbart, 2012; Weitz and Wilhelm, 2012; Jover *et al.*, 2014). Moreover, viruses constitute perhaps the biggest reservoir of genetic diversity in the oceans, increasing the cellular genetic diversity by gene transfer and, therefore, impacting the genetic diversity of all marine microbial populations (Jiang and Paul, 1998; Suttle, 2005).

For all these reasons viruses are an essential component of the marine ecosystem that have to be studied if we want to understand the ecology of the ocean and how it functions.

Virus–host interactions

Knowing the individual sequence, composition, abundance and/or behavior of a specific virus is important, but it is not enough to understand its biological role and impact in the environment. The first necessary step is to know who infects whom, i.e. relating a virus with its host. However, this is still a pending subject, since only a tiny fraction of the viruses known to date have an identified host (Not *et al.*, 2009; Sieradzki *et al.*, 2019).

All living organisms in the ocean are impacted by viral infections, from bacteria to protist and fish (Fig. 4). But viruses are not only involved in killing their host, and there are many different types of host–virus interactions that may have implications on the phylogeny and evolution of different components of the marine ecosystem (Middelboe and Brussaard, 2017). For example, infection by temperate viruses can prevent infection by similar viruses (called "superinfection exclusion mechanisms"), and contribute with important genetic information to the host, which can lead to genetic cellular evolution. Also, proviruses can make the cell less susceptible to be predated by grazers (Brüssow, 2007; Paul, 2008) and proviruses-encoded genes can contribute to the host functional properties, including virulence, by the so-called "lysogenic conversion", potentially expanding the niches occupied by the lysogenized hosts. Nucleocytoplasmic large DNA viruses (NCLDV) affect the mortality and diversity

of phytoplankton, and photosynthetic protists have been seen to produce several mechanisms of resistance to viruses as defense against NCLDV (e.g. Frada *et al.*, 2008; Van Etten *et al.*, 2010; Rolland *et al.*, 2019). Viruses may also acquire metabolic accessory genes from their hosts, like photosynthesis (Lindell *et al.*, 2004) or nutrient acquisition genes (Monier *et al.*, 2017), which expressed during infection may increase the fitness of the host, and ultimately increase virus production. Moreover, viruses may also control the expression of host genes during infection to promote viral production or inhibit host defense systems (Fig. 5) (Middelboe and Brussaard, 2017). Another important role in the host metabolism is that viruses interfere in their metabolic speediness (Sandaa, 2008). Also, some studies have revealed that the co-existence of bacteria competing for the same nutrients could be sustained by the viral lysis, limiting the number of each bacterial population (Bonachela and Levin, 2014). All these examples are only a small representation of the virus–host interactions that can happen in the marine ecosystem but exemplify why it is so important to study them.

With the exception of a few studies, most of the interactions studied until now have focused on phage–bacteria systems (e.g. Allers *et al.*, 2013; Labonté *et al.*, 2015), and much less is known about marine virus–eukaryote relationships. Therefore, there is a need to increase the studies encompassing eukaryotic systems to expand our knowledge in the field of marine virus–host ecology.

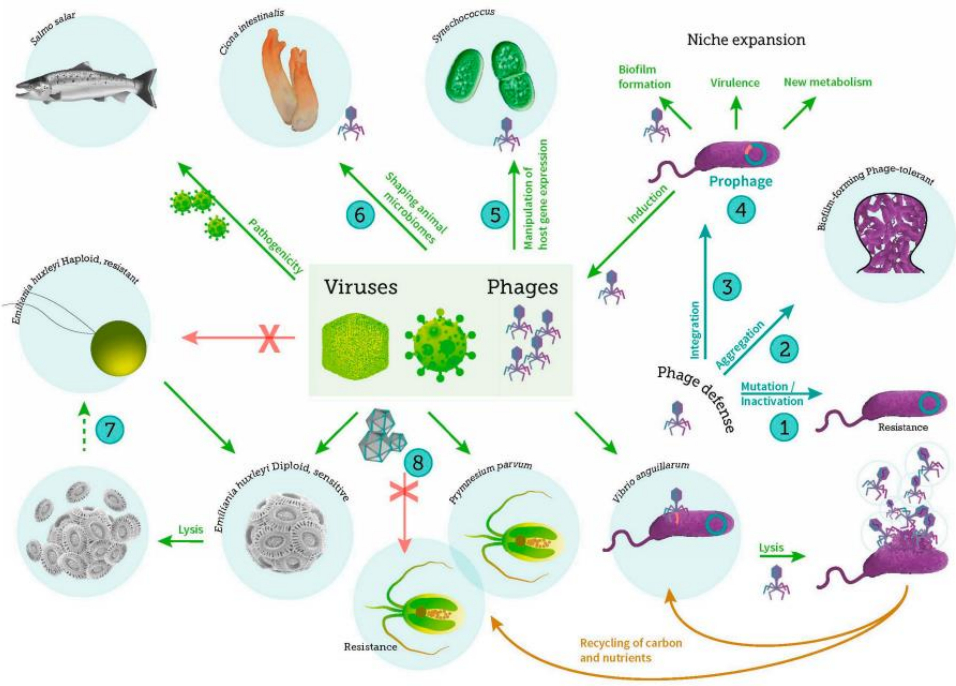


Figure 5. Schematic overview of some virus–host interactions in the marine ecosystems. (1) Cellular modification or enzymatic degradation of the incoming viral DNA to prevent viral infection; (2) Cellular aggregation or biofilm formation as defense against viruses; (3) Temperate viruses integration in the cellular genome can prevent infection by similar viruses; (4) Viruses may expand its metabolic or virulence properties contributing with important genetic information to the host; (5) Viruses can manipulate host gene expression to improve infection efficiency; (6) Phage interaction with their bacterial hosts contributes to shaping the gut microbiome of invertebrates; (7) in some phytoplankton, the diploid virally infected cells may undergo viral induced lysis or re-emerge as haploid cells containing viral RNA and lipids; (8) Nucleocytoplasmic large DNA viruses (NCLDV) infects a range of photosynthetic protists affecting mortality, diversity and production of phytoplankton (diagram from Middelboe and Brussaard, 2017).

How to detect viruses and their hosts: old and new approaches

At the end of the 20th century the first techniques to enumerate viruses were described. These traditional methods comprised different approaches, from the culture-based methods e.g. plaque counts and most-probable-number assays (Suttle and Chen, 1992), to the microscopy methods e.g. transmission electron microscopy (TEM) (Bergh *et al.*, 1989; Malenovska, 2013), epifluorescence microscopy combined with fluorescent staining of the viruses (Hennes *et al.*, 1995; Noble and Fuhrman, 1998), or flow cytometry (Marie *et al.*, 1999; Brussaard *et al.*, 2004). Culture-based methods are the only ones that allow discerning between infectious and non-infectious viruses, but they have the downside that they are constrained to cultivable hosts and their specific viruses. Microscopy methods also have some limitations: transmission electronic microscopy (TEM) (Fig. 1B) is a time-consuming method that requires expensive material and equipment, and epifluorescence microscopy, although it is relatively cheaper and renders virus detection accessible to field analysis, its magnification is limited to 1000x (Fig. 1A). Flow cytometry, is a sensitive and faster technique than TEM and epifluorescence microscopy, but as well as these two, it does not enable the distinction between infectious and non-infectious viruses and, moreover, the staining approach with SYBRGreen is biased towards dsDNA viruses (Martínez *et al.*, 2014). Therefore, we can approach the viral abundance, but we cannot identify the viruses or determine their host.

More recently several new methods have been described to detect, identify or enumerate different viruses from cultures and sea water. Depending on our system we can choose among several approaches. Some of the newest techniques are described in Table 1.

Table 1. Description, application, advantages and disadvantages of some common and emerging viral techniques for virus quantification and host identification (adapted from Breitbart *et al.*, 2018).

Method	Description	Application	Advantages	Disadvantages	Publications
Microfluidic digital PCR	Co-localization of host and viral DNA in microfluidic chambers by a simultaneous PCR amplification	Diversity; host identification	Viral–host linkage without culturing	Specialized equipment and sequence knowledge of virus and host to design the primers and probes	(Tadmor *et al.*, 2011)
PhageFISH	Co-localization of the virus and its host using FISH (Fluorescent in situ hybridization)	Quantification; diversity; potential host identification	Quantifies infected cells; potential for virus–host identification without culturing	Requires previous sequence knowledge of virus and host to design the primers and probes	(Allers *et al.*, 2013)
Single-cell genomics	Mining viral signals from the sequenced DNA genome of single cells	Diversity; host identification	Viral–host linkage without culturing	Database limitations	(Labonté *et al.*, 2015)
Metagenomics	Viral metagenomics (shotgun sequencing from viral communities) or mining viral signals from cellular metagenomes	Quantification; diversity	No need for known sequences; recovers higher diversity of viruses; sequences can be reanalyzed; provides ecological insights	Database limitations; computational/ bioinformatic limitations; difficulties on determining if the genetic material is of viral origin	(Roux *et al.*, 2017)
Polonies	Solid-phase PCR amplification that allows the simultaneous amplification of the host and its virus using degenerate primers	Quantification; diversity; potential host identification	Quantifies diverse viral groups using degenerate primers; potential for virus–host identification without culturing	Requires specialized equipment and previous sequence knowledge of virus and host to design the primers and probes	(Baran *et al.*, 2018)

As it is shown in Table 1, some of these techniques allow not only to detect viruses, their abundance, or their diversity, but also to identify their host/s

(mostly bacterial). Furthermore, they enable the study of both cultured and uncultured virus–host systems and their implications on mortality and dynamics of marine microbial communities (Middelboe and Brussaard, 2017).

However, all the approaches described in Table 1 have been basically applied to prokaryotic systems and, again, eukaryotes are the forgotten piece in marine ecology. Thus, we need to focus the attention on eukaryotic virus–host systems.

The eukaryotic virus–host systems

Protists are unicellular eukaryotes which are generally divided into photo- or heterotrophs depending on the source of carbon they use, although there is increasing evidence of mixotrophic protists, which have the capacity to acquire carbon both auto- and heterotrophically (Faure *et al.*, 2019). Phototrophic protists are main representatives of the phytoplankton community, which constitute the base of the marine microbial food web, releasing organic compounds to the environment and fueling bacterial growth (Jasti *et al.*, 2005). Heterotrophic and mixotrophic protists are grazers of virus, bacteria and other picoeukaryotes, and are trophic linkers and nutrient remineralizers (Bettarel *et al.*, 2005). Thus, protists play a crucial role in the epipelagic microbial food webs of the ocean (Gonzalez and Suttle, 1993; Sherr *et al.*, 1997; Fuhrman, 1999).

There is a wide spectrum of viruses that could infect them, from small RNA viruses as for example the Picorna-like viruses (Steward *et al.*, 2013), to giant DNA viruses as for example the giant Mimivirus (Raoult, 2004). The viral infection of photoautotrophic and heterotrophic microorganisms makes that an important fraction of cellular carbon returns to the water column as dissolved organic matter (DOC), influencing the particle size-distribution, nutrient cycling and biological activity of the ecosystem (Suttle, 2007; Coy *et al.*, 2018). Hence, the study of virus–protist systems is crucial due to all the implications that they have in the functioning of the marine trophic web.

Cultured vs uncultured eukaryotic hosts

It is widely known that only a few species of bacteria have culture representatives (Zengler *et al.*, 2002; Joint *et al.*, 2010), and less than ~1% of the bacteria on Earth can be easily cultivated (Vartoukian *et al.*, 2010). In the case of protists, these numbers are still much lower.

The protist culturing bias can be explained by the isolation media, which drives a shift in the community composition to favor certain species. For example, bacterivorous heterotrophic protists are typically cultivated using seawater supplemented with cereals, rice or yeast that promote the growth of bacteria as food. Nevertheless, this rich media will fuel the growth of some large and abundant bacteria, which in turn will trigger the growth of only the pool of protist species that can feed on them. Typically, this pool of protist corresponds to the rare ones in natural occurring communities based on culture-independent approaches (Jürgens and Massana, 2008; del Campo and Massana, 2011). Therefore, some of the most abundant and representative heterotrophic protists in the marine environment refuse cultivation, as it happens with bacteria.

However, there are some model organisms that are amenable to culture and that constitute the perfect system to study protist–virus interactions, and test and implement approaches that can be later applied in nature. One of these approaches is the fluorescent *in situ* hybridization (FISH). Since FISH appeared several variations of the method have been developed. One of these variants is phageFISH (Table 1), which allows to visually detect the interaction between a phage and its host. This approach presents a high potential to be extended to other systems (eukaryotic) and to the environment.

Another important bias in the field is that most eukaryotic genomes available are focused on multicellular eukaryotes and their parasites. More than 96% of the described eukaryotic species are Metazoa, Fungi or Embryophyta (del Campo *et al.*, 2014). However, when we focus on protists we see that most of

the species present in culture collections and/or genome projects are phototrophic species or economically important cells (del Campo *et al.*, 2014), therefore, heterotrophic protists are in general less studied. Since few years ago some culture-independent approaches have been developed to study uncultured cells and overcome these limitations. One of the most powerful tools to study both photo- and heterotrophic protists is the single cell genomics (SCG) approach (Table 1). SCG is the perfect complement to cultivation providing genomic information from individual uncultured cells (Stepanauskas, 2012). It also offers the opportunity to study uncultured virus–host interactions and may allow increasing the eukaryotic viral community databases (e.g. Labonté *et al.*, 2015; Roux *et al.*, 2016).

Therefore, the combination of both culture and culture-independent methodologies can improve our understanding of the ecology of marine protists and their interactions with viruses.

AIM OF THE THESIS

The main goal of this thesis is to study virus–host interactions in marine picoeukaryotic cells at the single cell level. To achieve it, the dissertation contains three chapters that are structured in the following main objectives:

1) **To implement the VirusFISH technique to visualize viral infection dynamics, as well as to quantify free viral production, in a model culture system.**

This objective was addressed using the model system *Ostreococcus tauri* and its virus OtV5, and the results are compiled in Chapter 1, which includes:

- Detection and monitoring of the induced infection of *Ostreococcus tauri* with the virus OtV5 by VirusFISH.
- Determination of the abundance of free OtV5 particles produced during the infection in the non-axenic culture.
- Calculation of *O. tauri* burst size.

2) **To demonstrate the validity of VirusFISH to investigate populations-specific virus–host dynamics in nature.**

To accomplish this objective we applied VirusFISH to visualize the seasonal dynamics of *Ostreococcus* spp. viral infection during an annual cycle in the Cantabrian Sea. These results are compiled in Chapter 2, which includes:

- The study of *Ostreococcus* spp. abundance and the interaction with their viruses, with the quantification of the impact of viruses on *Ostreococcus* populations.
- Comparison of VirusFISH results with transcriptional activities of virus and hosts derived from metatranscriptomic data from the same samples.

3) To assess the viral content of uncultured prevalent marine picoeukaryotes from the global ocean using single-cell genomics.

To achieve this goal we used genomic approaches to detect virus–host interactions in single amplified genomes of uncultured picoeukaryotes collected during the *Tara* Oceans expedition. These results are compiled in chapter 3 of this thesis, which includes:

- Identification of viral signals in single amplified genomes of 64 Stramenopiles using genomic approaches.
- Studying the biogeographic distribution of the identified viral sequences using global ocean metagenomes.
- Extensive analysis of viral sequences, detected in four SAGs, extremely close to the virophage *mavirus*.

Each chapter is structured as a scientific paper to facilitate the comprehension of the thesis. Chapter 1 is currently under review in Applied and Environmental Microbiology (preprint in bioRxiv, doi: 10.1101/849455). Chapter 2 will soon be submitted. Chapter 3 was published in September 2019 in the Molecular Ecology journal (doi:10.1111/mec.15210). The state of the art and specific methodologies are presented within each chapter, with a discussion of the results obtained.

REFERENCES

Abedon, S.T. (2008) Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses Abedon,S.T. (ed) Cambridge University Press, Cambridge.

Allers, E., Moraru, C., Duhaime, M.B., Beneze, E., Solonenko, N., Barrero-Canosa, J., et al. (2013) Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environ. Microbiol.* **15**: 2306–2318.

Aloia, R.C., Tian, H., and Jensen, F.C. (1993) Lipid composition and fluidity of the human immunodeficiency virus envelope and host cell plasma membranes. *Proc. Natl. Acad. Sci.* **90**: 5181–5185.

Baran, N., Goldin, S., Maidanik, I., and Lindell, D. (2018) Quantification of diverse virus populations in the environment using the polony method. *Nat. Microbiol.* **3**: 62–72.

Bergh, Ø., BØrsheim, K.Y., Bratbak, G., and Heldal, M. (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467–468.

Bettarel, Y., Sime-Ngando, T., Bouvy, M., Arfi, R., and Amblard, C. (2005) Low consumption of virus-sized particles by heterotrophic nanoflagellates in two lakes of the French Massif Central. *Aquat. Microb. Ecol.* **39**: 205–209.

Bonachela, J.A. and Levin, S.A. (2014) Evolutionary comparison between viral lysis rate and latent period. *J. Theor. Biol.* **345**: 32–42.

Breitbart, M. (2012) Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**: 425–448.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018) Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**: 754–766.

Brussaard, C.P.., Noordeloos, A.A.., Sandaa, R.-A., Heldal, M., and Bratbak, G. (2004) Discovery of a dsRNA virus infecting the marine photosynthetic protist Micromonas pusilla. *Virology* **319**: 280–291.

Brussaard, C.P.D., Wilhelm, S.W., Thingstad, F., Weinbauer, M.G., Bratbak, G., Heldal, M., et al. (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J.* **2**: 575–578.

Brüssow, H. (2007) Bacteria between protists and phages: from antipredation strategies to the evolution of pathogenicity. *Mol. Microbiol.* **65**: 583–589.

Bzhalava, Z., Hultin, E., and Dillner, J. (2018) Extension of the viral ecology in humans using viral profile hidden Markov models. *PLoS One* **13**: e0190938.

del Campo, J. and Massana, R. (2011) Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. *Protist* **162**: 435–448.

del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**: 252–259.

Cochlan, W.P., Wikner, J., Steward, G.F., Smith, D.C., and Azam, F. (1993) Spatial distribution of viruses, bacteria and chlorophyll a in neritic, oceanic and estuarine environments. *Mar. Ecol. Prog. Ser.* **92**: 77–87.

Coy, S., Gann, E., Pound, H., Short, S., and Wilhelm, S. (2018) Viruses of

eukaryotic algae: diversity, methods for detection, and future directions. *Viruses* **10**: 487.

Daniels, R., Sadowicz, D., and Hebert, D.N. (2007) A very late viral protein triggers the lytic release of SV40. *PLoS Pathog.* **3**: e98.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J.A., Middelburg, J.J., Noble, R.T., and Suttle, C.A. (2011) Marine viruses and global climate change. *FEMS Microbiol. Rev.* **35**: 993–1034.

Echols, H. (1972) Developmental pathways for the temperate phage: lysis vs lysogeny. *Annu. Rev. Genet.* **6**: 157–190.

Van Etten, J.L., Lane, L.C., and Dunigan, D.D. (2010) DNA viruses: the really big ones (giruses). *Annu. Rev. Microbiol.* **64**: 83–99.

Faure, E., Not, F., Benoiston, A.-S., Labadie, K., Bittner, L., and Ayata, S.-D. (2019) Mixotrophic protists display contrasted biogeographies in the global ocean. *ISME J.* **13**: 1072–1083.

Filée, J. (2018) Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr. Opin. Virol.* **33**: 81–88.

Frada, M., Probert, I., Allen, M.J., Wilson, W.H., and de Vargas, C. (2008) The "Cheshire Cat" escape strategy of the coccolithophore *Emiliania huxleyi* in response to viral infection. *Proc. Natl. Acad. Sci.* **105**: 15944–15949.

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Gonzalez, J.M. and Suttle, C.A. (1993) Grazing by marine nanoflagellates on viruses and virus-sized particles: ingestion and digestion. *Mar. Ecol. Prog. Ser.* **94**: 1-10.

Goodwin, C.M., Xu, S., and Munger, J. (2015) Stealing the keys to the kitchen: viral manipulation of the host cell metabolic network. *Trends Microbiol.* **23**: 789–798.

Grayson, P. and Molineux, I.J. (2007) Is phage DNA 'injected' into cells—biologists and physicists can agree. *Curr. Opin. Microbiol.* **10**: 401–409.

Hennes, K.P., Suttle, C.A., and Chan, A.M. (1995) Fluorescently labeled virus probes show that natural virus populations can control the structure of marine microbial communities. *Appl. Environ. Microbiol.* **61**: 3623–3627.

Jasti, S., Sieracki, M.E., Poulton, N.J., Giewat, M.W., and Rooney-Varga, J.N. (2005) Phylogenetic diversity and specificity of bacteria closely associated with *Alexandrium* spp. and other phytoplankton. *Appl. Environ. Microbiol.* **71**: 3483–3494.

Jiang, S.C. and Paul, J.H. (1998) Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64**: 2780–7.

Joint, I., Mühling, M., and Querellou, J. (2010) Culturing marine bacteria - an essential prerequisite for biodiscovery. *Microb. Biotechnol.* **3**: 564–575.

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

Jürgens, K. and Massana, R. (2008) Protistan Grazing on Marine Bacterioplankton. In, *Microbial Ecology of the Oceans*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 383–441.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* **9**: 2386–2399.

Lara, E., Vaqué, D., Sà, E.L., Boras, J.A., Gomes, A., Borrull, E., et al. (2017) Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci. Adv.* **3**: e1602565.

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

Malenovska, H. (2013) Virus quantitation by transmission electron microscopy, TCID50, and the role of timing virus harvesting: A case study of three animal viruses. *J. Virol. Methods* **191**: 136–140.

Marciano, D.K. (1999) An Aqueous Channel for Filamentous Phage Export. *Science* **284**: 1516–1519.

Marie, Brussaard, Thyrhaug, Bratbak, and Vaulot (1999) Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**: 45–52.

Martínez, J.M., Swan, B.K., and Wilson, W.H. (2014) Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J.* **8**: 1079–1088.

Maynard, N.D., Gutschow, M. V, Birch, E.W., and Covert, M.W. (2010) The virus as metabolic engineer. *Biotechnol. J.* **5**: 686–694.

Middelboe, M. and Brussaard, C. (2017) Marine viruses: key players in marine ecosystems. *Viruses* **9**: 302.

Monier, A., Chambouvet, A., Milner, D.S., Attah, V., Terrado, R., Lovejoy, C., et al. (2017) Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proc. Natl. Acad. Sci.* **114**: E7489–E7498.

Munn, C.B. (2006) Viruses as pathogens of marine organisms—from bacteria to whales. *J. Mar. Biol. Assoc. UK* **86**: 453–467.

Noble, R. and Fuhrman, J. (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**: 113–118.

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS One* **4**: e7143.

Paul, J.H. (2008) Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* **2**: 579–589.

Pernice, M.C., Forn, I., Gomes, A., Lara, E., Alonso-Sáez, L., Arrieta, J.M., et al. (2015) Global abundance of planktonic heterotrophic protists in the deep ocean. *ISME J.* **9**: 782–792.

Pimentel, M. (2014) Genetics of Phage Lysis. *Microbiol. Spectr.* **2**: MGM2-0017–2013.

Raoult, D. (2004) The 1.2-Megabase genome sequence of Mimivirus. *Science* **306**: 1344–1350.

Rolland, C., Andreani, J., Louazani, A., Aherfi, S., Francis, R., Rodrigues, R., et al. (2019) Discovery and further studies on giant viruses at the IHU Mediterranee infection that modified the perception of the virosphere. *Viruses* **11**: 312.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., et al.

(2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689–693.

Roux, S., Chan, L.-K., Egan, R., Malmstrom, R.R., McMahon, K.D., and Sullivan, M.B. (2017) Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**: 858.

Sandaa, R.-A. (2008) Burden or benefit? Virus–host interactions in the marine environment. *Res. Microbiol.* **159**: 374–381.

Saussereau, E. and Debarbieux, L. (2012) Bacteriophages in the experimental treatment of *Pseudomonas aeruginosa* infections in mice. In, *Advances in Virus Research*. Academic Press, pp. 123–141.

Schrad, J., Young, E., Abrahão, J., Cortines, J., and Parent, K. (2017) Microscopic characterization of the brazilian giant Samba virus. *Viruses* **9**: 30.

Sherr, E.B., Sherr, B.F., and Fessenden, L. (1997) Heterotrophic protists in the Central Arctic Ocean. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **44**: 1665–1682.

Sieradzki, E.T., Ignacio-Espinoza, J.C., Needham, D.M., Fichot, E.B., and Fuhrman, J.A. (2019) Dynamic marine viral infections and major contribution to photosynthetic processes shown by spatiotemporal picoplankton metatranscriptomes. *Nat. Commun.* **10**: 1169.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**: 613–620.

Steward, G.F., Culley, A.I., Mueller, J.A., Wood-Charlson, E.M., Belcaid, M., and Poisson, G. (2013) Are we missing half of the viruses in the ocean? *ISME J.* **7**: 672–679.

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

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

Suttle, C.A. (1994) The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol* **28**: 237–243.

Suttle, C.A. and Chen, F. (1992) Mechanisms and rates of decay of marine viruses in seawater. *Appl. Environ. Microbiol.* **58**: 3721–9.

Tadmor, A.D., Ottesen, E.A., Leadbetter, J.R., and Phillips, R. (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**: 58–62.

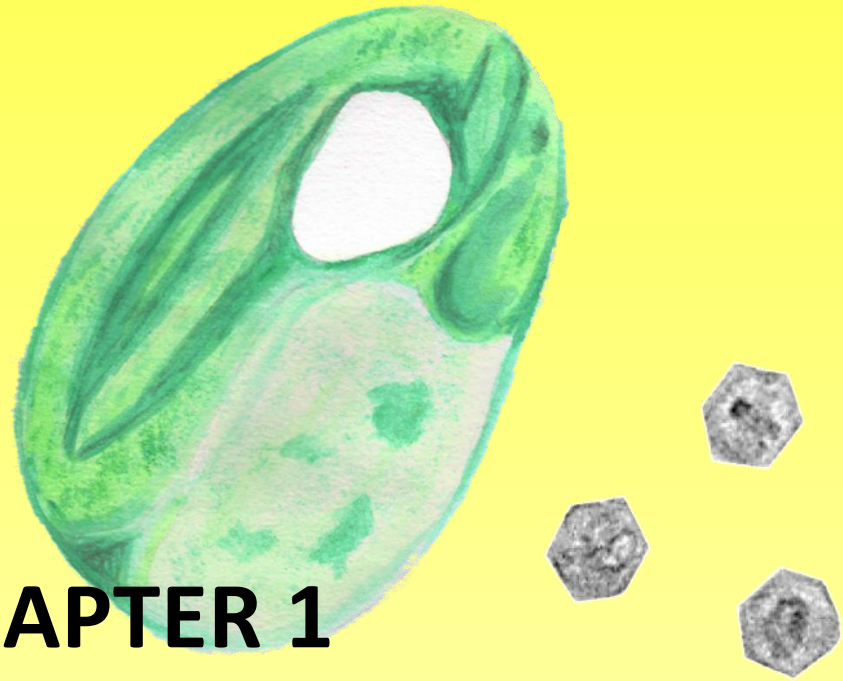
Vartoukian, S.R., Palmer, R.M., and Wade, W.G. (2010) Strategies for culture of 'unculturable' bacteria. *FEMS Microbiol. Lett.* **309**: 1–7.

Weitz, J. and Wilhelm, S. (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**: 17.

Yamauchi, Y. and Helenius, A. (2013) Virus entry at a glance. *J. Cell Sci.* **126**: 1289–1295.

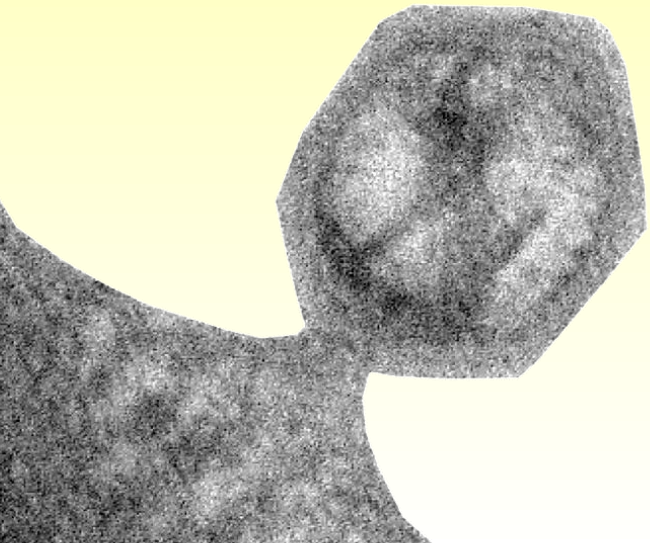
York, A. (2017) A bacteriophage-like entry pathway in eukaryotes. *Nat. Rev. Microbiol.* **15**: 577–577.

Zengler, K., Toledo, G., Rappe, M., Elkins, J., Mathur, E.J., Short, J.M., and Keller, M. (2002) Cultivating the uncultured. *Proc. Natl. Acad. Sci.* **99**: 15681–15686.



CHAPTER 1

Visualization of viral infection dynamics in a unicellular eukaryote and quantification of viral production using VirusFISH



Castillo, Y. M., Sebastián, M., Forn, I., Grimsley, N., Yau, S., Moraru, C. and Vaqué, D. Visualization of viral infection dynamics in a unicellular eukaryote and quantification of viral production using VirusFISH. Submitted to Applied and Environmental Microbiology. Preprint available in bioRxiv, doi: 10.1101/849455

ABSTRACT

One of the major challenges in viral ecology is to assess the impact of viruses in controlling the abundance of specific hosts in the environment. For this, techniques that enable the detection and quantification of virus–host interactions at the single-cell level are essential. With this goal in mind, we implemented VirusFISH (Virus Fluorescence *in situ* Hybridization) using as a model the marine picoeukaryote *Ostreococcus tauri* and its virus OtV5. VirusFISH allowed the visualization and quantification of the fraction of infected cells during an infection experiment. We were also able to quantify the abundance of free viruses released during cell lysis and assess the burst size of our non-axenic culture, because we could discriminate OtV5 from phages. Our results showed that although the major lysis of the culture occurred between 24 and 48 h after OtV5 inoculation, some new viruses were produced between 8 and 24 h, propagating the infection. Nevertheless, the production of viral particles increased drastically after 24 h. The burst size for the *O. tauri*–OtV5 system was 7±0.4 OtV5 per cell, which was consistent with the estimated amount of viruses inside the cell prior to cell lysis. With this work we demonstrate that VirusFISH is a promising technique to study specific virus–host interactions in non-axenic cultures, and set the ground for its application in complex natural communities.

KEYWORDS

VirusFISH; *Ostreococcus tauri*; OtV5; virus–host interactions; culture system; marine picoeukaryote.

1.1. INTRODUCTION

Marine viruses have been studied during the last three decades, mostly by traditional approaches as microscopy (Noble and Fuhrman, 1998) and flow cytometry (Marie *et al.*, 1999), used for the enumeration and estimation of viral production. However, in the last few years, the development of high throughput sequencing techniques has considerably changed the field, and our knowledge about viral communities has exponentially increased. These new sequencing approaches provide information about the viral taxonomic and genomic diversity, about their biogeography and, to a certain extent, about their potential hosts (*e.g.* Chow *et al.*, 2015; Labonté *et al.*, 2015). However, they do not allow the visualization of specific virus–host interactions and the monitoring of the infection dynamics, which are crucial to better understand the contribution of viruses in shaping microbial communities and biogeochemical cycles.

Attempts to identify virus–host associations date back to the 90s, when the role of viruses in the marine environment started to be recognized. Hennes *et al.* (1995) pioneered an approach to identify and enumerate specific virus-infected bacteria in natural communities by using fluorescently stained viruses (labeled with YOYO-1 or POPO-1) as probes and epifluorescence microscopy. Years after, Tadmor *et al.* (2011) used microfluidic digital PCR to detect specific phage–host associations in the termite gut. With this method, they managed to directly detect the phage–host association by targeting genes from both components without culturing, but with no visual representation of the infection.

A few years ago, Allers *et al.* (2013) developed phageFISH and used it to monitor phage infections at the single-cell level in a marine podovirus–gammaproteobacterial host system. PhageFISH uses mixtures of polynucleotide probes labeled with digoxigenin to target phage genes, and a single HRP labeled oligonucleotide probe to target host rRNA. The signal from the two types of probes is amplified and visualized by catalyzed reporter

deposition (CARD) of fluorescently labeled tyramides. Compared to the method from Hennes *et al.,* (1995), where the infection was forced by adding stained viruses to identify the host within natural communities, phageFISH enables the visualization of the infection dynamics of specific virus–host pairs, because it simultaneously targets the virus and the host. More recently developed, direct-geneFISH (Barrero-Canosa *et al.*, 2017) uses simultaneously a mixture of polynucleotide probes directly labeled with fluorochromes, to detect specific genes in cells, and a single oligonucleotide probe, carrying multiple fluorochromes, to identify bacterial cells.

In the present work, based on phageFISH and direct-geneFISH, we developed the VirusFISH technique with the aim to allow i) identification and quantification of specific virus–unicellular eukaryote interactions at the single-cell level and ii) identification and quantification of free virus particles. VirusFISH consists of two steps. First, a CARD-FISH step is used to detect host cells, with HRP-labeled oligonucleotide probes targeting the 18S rRNA. Then, a VirusFISH step is applied to detect viruses, using multiple polynucleotide probes directly labeled with fluorochromes that target viral genes. VirusFISH can be used to detect both intracellular viruses and free viral particles.

As proof of principle, we used VirusFISH to monitor viral infections of the unicellular green alga *Ostreococcus tauri* (*O. tauri*), the smallest known marine photosynthetic eukaryote, with the virus *Ostreococcus tauri* virus 5 (OtV5).

1.2. MATERIALS AND METHODS

1.2.1. Experimental viral infection of *O. tauri*

The host strain *O. tauri* RCC4221 (Roscoff Culture Collection, NCBI accession number txid70448) was grown in 60 mL of L1 medium (Guillard and Hargraves, 1993) in aerated flasks (Sarstedt), and incubated at 21.5°C (±0.5°C) with white light ~100 µE and a 10:14 hours photoperiod (light:darkness), until stationary

phase (7.16×10^7 ± 3.57×10^6 cells mL^{-1}, estimated by 4'-6-Diamidino-2-phenylindole (DAPI) counts). Triplicate *O. tauri* cultures (20 mL) were infected with 1 mL of OtV5 inoculum (1.3×10^7 ± 4.3×10^6 viruses mL^{-1}, estimated by plaque-forming units), resulting in a 0.01 MOI (multiplicity of infection). Non-infected triplicate *O. tauri* cultures (inoculated with 1 mL of L1 medium) were used as control. After OtV5 inoculation, samples (900 µL) were taken over 3 days at times 0, 8, 24, 48 and 72 h, and fixed with 100 µL of freshly filtered formaldehyde (3.7% final concentration) for 15 min at room temperature. Then, 500 µL of fixed sample were filtered through 0.2 µm pore size polycarbonate white filters (Merck™ GTTP02500) to retain cells, and through 0.02 µm pore size anodisc filters (Whatman®) (after a 0.2 µm pore size prefiltration to remove cells and debris) to retain free viruses. Polycarbonate filters of 0.2 µm pore size were embedded in 0.1% (w v^{-1}) low gelling point agarose and treated for 1h with 96% ethanol and 1h with pure methanol, to remove cellular pigments that can interfere with the CARD-FISH signal (Fig. S1), and 10 min with HCl to inactivate endogenous peroxidases (Pavlekovic *et al.*, 2009). All filters were kept at -20°C until hybridization.

1.2.2. OtV5 probe design and synthesis

For the detection of the OtV5 virus (NCBI accession number EU304328) we designed 11 dsDNA polynucleotide probes (300 bp each) using the software geneProber web service (http://gene-prober.icbm.de/). These 11 probes covered a total of 3998 bp of the OtV5 viral genome, offering sufficient sensitivity to detect single viruses (Table S1), as it has been shown before (Barrero-Canosa *et al.*, 2017). Each probe synthesis was done by obtaining the corresponding polynucleotides by PCR, and then all probes were mixed and labeled with the Alexa594 fluorochrome, based on the protocol from Barrero-Canosa *et al.* (2017). The PCR was set up as follows: 10pg of OtV5 DNA were added to a reaction mixture containing 200 µM (each) deoxyribonucleoside triphosphates (Invitrogen), 1 µM of each primer, 1x PCR buffer (Invitrogen), and 5U of *Taq* DNA polymerase (Invitrogen). The thermal cycling was performed in a

C1000TM Thermal Cycler (Bio-Rad) with an initial denaturation step at 95°C (5 min), followed by 30 rounds at 95°C (1 min), X°C (30 s), and 72°C (30 s), and a final extension at 72°C (10 min). X value corresponds to the optimal annealing temperature for each of the primers, determined after performing gradient PCRs. All OtV5 primers had an optimal annealing temperature of 62.5°C, with the exception of primers #3 and #5 that had an annealing temperature of 65.5°C. Primers sequences can be found in Table S1. For each polynucleotide, several PCRs were done to obtain a minimum of 400µL PCR reaction volume. This volume was purified on a single purification column using the QIAquick PCR purification kit – Qiagen, cat.no. 28106, and resuspended in a TE solution (5 mM Tris-HCl, 1 mM EDTA, pH 8.0). The polynucleotide length was checked by agarose gel electrophoresis, and the concentration was measured spectrophotometrically using a NanoDrop 1000 (Fisher Thermo Scientific). Further, all 11 polynucleotides were mixed equimolarly to yield a total of 1 µg DNA in 10 µl TE. Later, the probe mixture was heated to 95°C for 5 min to denature it and then incubated for 30 min at 80°C with 10 µL of the red emission dye Alexa594 (Ulysis™ Alexa Fluor® 594 Nucleic Acid Labeling Kit, Thermofisher, cat.no: U21654). The unbound Alexa594 was removed using chromatography columns (Micro Bio-spin chromatography columns P-30, Bio-Rad, cat.no. 732-6202). The concentration of the probe mixture and the labeling efficiency with Alexa 594 were determined spectrophotometrically using a NanoDrop 1000 with the Multi-Array option and N-50. For a successful detection of the virus, we observed that the labeling efficiency should be higher than 6 Alexas per probe. Fluorescent probes were stored at -20°C until use.

1.2.3. Detection of *O. tauri* cells using 18S rRNA targeted CARD-FISH

O. tauri cells were labeled using Catalyzed Reporter Deposition-FISH (CARD)-FISH (Pernice *et al.*, 2015) with the 18S rRNA targeted probe OSTREO01 for *Ostreococcus* spp. (Not *et al.*, 2004). Briefly, the hybridization was carried out by covering filter pieces with 20 µL of hybridization buffer with 40% deionized formamide and incubating at 35°C overnight. After two successive washing steps

of 10 min at 37°C in a washing buffer and a equilibration in phosphate-buffered saline for 15 min at room temperature (Cabello *et al.*, 2016), the signal was amplified for 1h at 46°C with Alexa488-labeled tyramide. Filters were then placed in phosphate-buffered saline two times for 10 min, rinsed with MilliQ water and air-dried.

1.2.4. Detection of intracellular and free OtV5 viruses using VirusFISH

OtV5 viruses were labeled using VirusFISH, a modified version of the direct-geneFISH protocol (Barrero-Canosa *et al.*, 2017). VirusFISH was applied on i) 0.2 µm pore size filters that were previously hybridized with the CARD-FISH probes for the host to monitor the infection, ii) 0.02 µm pore size filters to monitor the dynamics of the free OtV5 viruses. The hybridization was done by covering the filter pieces with 25 µL of 40% formamide hybridization buffer (HB) containing OtV5 probes and incubating first for 40 minutes at 85°C, and then for 2 h at 46°C. The composition of the hybridization buffer was: 40% formamide, 5x saline-sodium citrate, 20% dextran sulfate, 0.1% sodium dodecyl sulfate, 20 mM EDTA, 0.25 mg mL^{-1} sheared salmon sperm, 0.25 mg mL^{-1} yeast RNA and 1% blocking reagent). The volume of probe mixture labeled with Alexa594 to add to the HB was calculated based on the following formula, according to Barrero-Canosa *et al.* (2017):

$$\frac{(25\mu L\ HB \cdot \text{number of filters}) \cdot (\frac{62pg}{\mu L}\ \text{final probe concentration} \cdot \text{number of total probes})}{\text{Viral probe concentration } \left(\frac{ng}{\mu L}\right) \cdot 1000} = \mu L \text{ probe mixture}$$

Assuming that the volume of HB for each filter portion is 25 µL and 62 pg μL^{-1} is the desired final probe concentration according to Barrero-Canosa et al. (2017).
Finally, samples were washed at 48°C for 15 minutes with gentle shaking in a washing buffer (560 µL NaCl 5M, 1mL Tris-HCl 1M pH8, 1mL EDTA 0.5M pH8 and 50µL 10% sodium dodecyl sulfate in 50mL of autoclaved MilliQ water), rinsed with MilliQ water and air-dried.

1.2.5. Sample mounting, visualization and image analysis

After hybridization, 0.2 µm filters were counterstained with DAPI at 0.5 µg mL^{-1} to observe *O. tauri* nuclei, and mounted in antifading reagent (77% glycerol, 15% VECTASHIELD and 8% 20x PBS) (Cabello *et al.*, 2016). Images were manually acquired using a Zeiss Axio Imager Z2m epifluorescence microscope (Carl Zeiss, Germany) connected to a Zeiss camera (AxioCamHR, Carl Zeiss MicroImaging, S.L., Barcelona, Spain) at x1000 magnification through the AxioVision 4.8 software. *O. tauri* was observed by epifluorescence microscopy under blue light (475/30 nm excitation, 527/54 BP emission, and FT 495 beam splitter) and OtV5 under orange light (585/35 nm excitation, 615 LP emission, and FT 570 beam splitter). All pictures were taken using the same intensities and exposure times (400 ms for the *O. tauri* and 1 s for the virus detection).

Total free viruses (i.e. both OtV5 and phages present in the non-axenic culture) collected on the 0.02µm pore size filters, were counterstained with SYBRGold (SYBR™ Gold solution, Invitrogen) at 2x final concentration for 12 min, and then rinsed abundantly with MilliQ water to remove excess staining. Filters were finally mounted on slides with an antifading mounting solution (CitiFluor™ Glycerol-PBS Solution AF1). Images were acquired on the same Zeiss microscope and camera at x1000 magnification. OtV5 were observed by epifluorescence microscopy under orange light (585/35 nm excitation, 615 LP emission, and FT 570 beam splitter) and total viruses (OtV5 and phages) under blue light (475/30 nm excitation, 527/54 BP emission, and FT 495 beam splitter). All pictures were taken using the same intensities and exposure times as mentioned above. Image analysis for free virus detection was done using the software ACMEtool 3 (July 2014; M Zeder, Technobiology GmbH, Buchrain, Switzerland).

During the image analysis we observed that a fraction of OtV5 virions released from the cells during lysis was trapped on the extracellular organic matrix around the cells (here referred to as viral clouds) (Weinbauer *et al.*, 2009), and retained on the 0.2 µm filters. Thus, for 48 and 72 hours, we calculated the viral

abundance of OtV5 retained on the 0.2 µm polycarbonate filters from the average area of viral clouds corrected by the average area of an OtV5 virus (48 h, n=2432 viral clouds areas; 72 h, n=307 viral clouds areas; OtV5, n=30,000 OtV5 particles areas). Consequently, we considered the total OtV5 production at 48 and 72 hours as the sum of the free virus abundance collected onto the 0.02 µm filters plus the viral abundance retained on the 0.2 µm. Areas were determined using the AxioVision 4.8 software (Schindelin *et al.*, 2012).

1.2.6. Burst size estimations

Burst size was calculated based on the formula established by Middelboe and Lyck, (2002) that compares the Δvirus abundance / Δhost abundance at the times when the host decline happens (here, between 24 and 48 hours). To corroborate the burst size by the classical method, we also assessed the area of the host occupied with OtV5 at late stages of the infection. Since Henderson *et al.* (2007) reported that the structure of *Ostreococcus* is rather flattened, we calculated the average area of *O. tauri* hosting the OtV5 virions at 24 h (n=90 areas), when the maximum infection was observed, and the average area of single free OtV5 particles (n=30,000 viruses). The capacity was finally estimated by dividing the average cellular area occupied with viruses by the average area of a single viral particle. Areas were determined using the AxioVision 4.8 software (Schindelin *et al.*, 2012).

1.3. RESULTS

1.3.1. The OtV5 – *O. tauri* infection dynamics as revealed by VirusFISH

A non-axenic culture of *O. tauri* was infected with the virus OtV5, at a MOI of 0.01 and an uninfected culture was grown in parallel, as a control (Fig. 1A). Using VirusFISH, the two cultures were followed for 72 h, quantifying i) the absolute abundance of *O. tauri* cells and ii) the relative and absolute abundance of infected *O. tauri* cells. The infected culture experienced a dramatic decrease in

cell density of two orders of magnitude between 24 and 48 h (Fig. 1B, Fig. 2 and Fig. S2). At 72h, almost no *O. tauri* cells were detected (Fig. S2), consistent with the clearing of the infected culture (Fig. 1A).

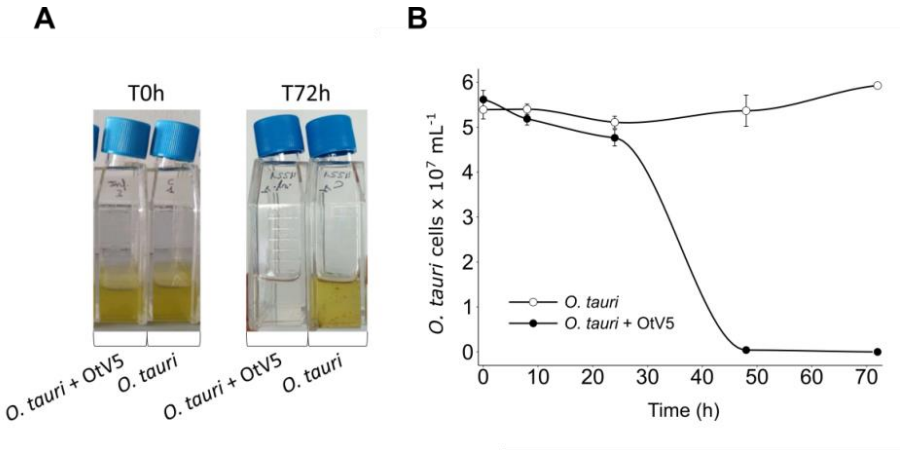


Figure 1. Dynamics of the infection of *Ostreococcus tauri* with OtV5. **A.** Infection and control culture flasks at time 0h and 72h. **B.** *O. tauri* CARD-FISH cell abundances (cells x 10^7 ml^{-1}) counted by epifluorescence microscopy in both the infected (solid circles) and the control (empty circles) triplicate cultures.

At the MOI used, rapid adsorption of all the viral particles added would theoretically result in 1% of infected cells. However, despite infected cells were visible as early as 0.4 h, the abundance was very low at both 0.4 and 8 h (0.02% and 0.2%, respectively), suggesting that not all viral particles had yet been adsorbed (Fig. 3). Nevertheless, the fact that at 24 h we found 16% of the population infected implies that new viruses had already been produced that had gone on to infect more cells in the culture. Later, at 48 h, the abundance of cells decreased by two orders of magnitude and 60% of the remaining cells were infected (Fig. 2 and 3). In contrast, the abundance of *O. tauri* cells in the control cultures remained relatively constant along the experiment and, as expected, no infected cells were observed (Fig. 1B and S3).

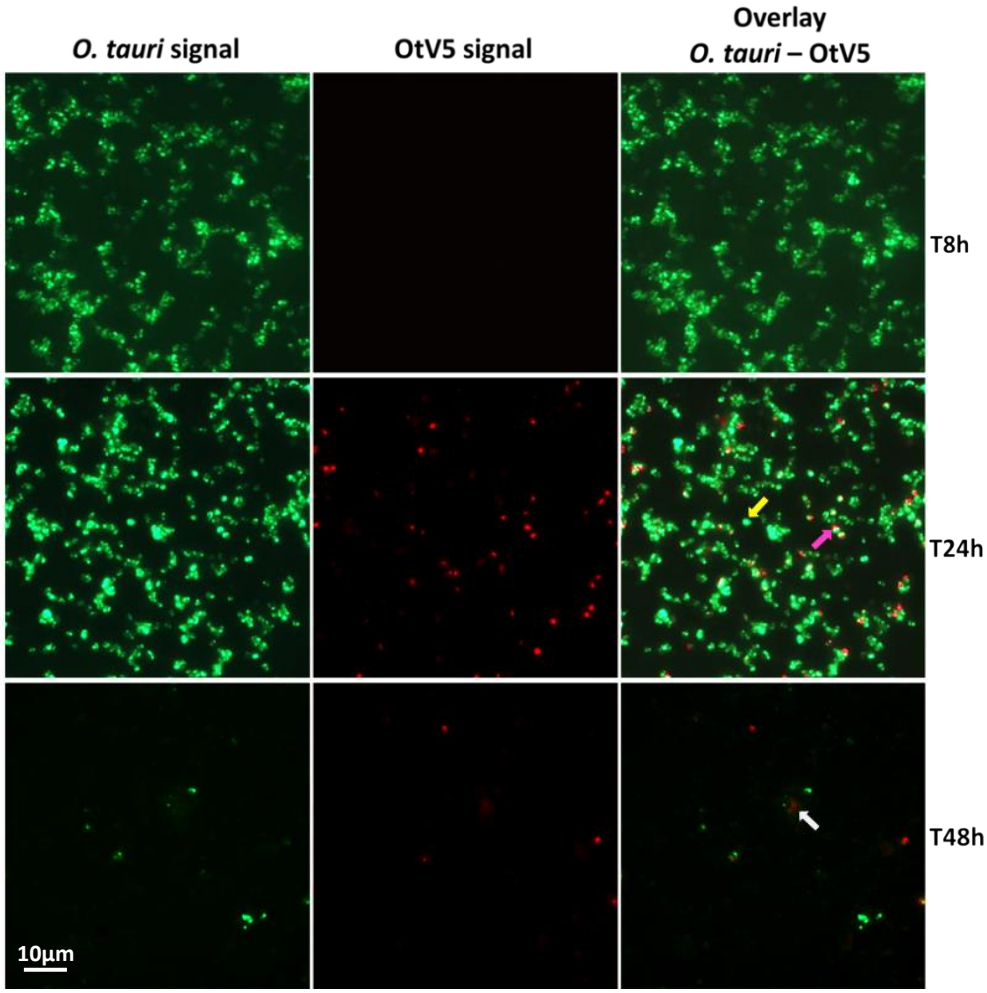


Figure 2. Micrographs of the evolution of the infection from time 8h to 48h. Left: *O. tauri* only. Centre: OtV5 only. Right column: overlay of *O. tauri* host cells in green (Alexa488) and virus in red (Alexa594). Yellow arrow: non-infected *O. tauri*; pink arrow: infected *O tauri*; grey arrow: cloud of viruses retained on the filter by the organic matter released during the lysis.

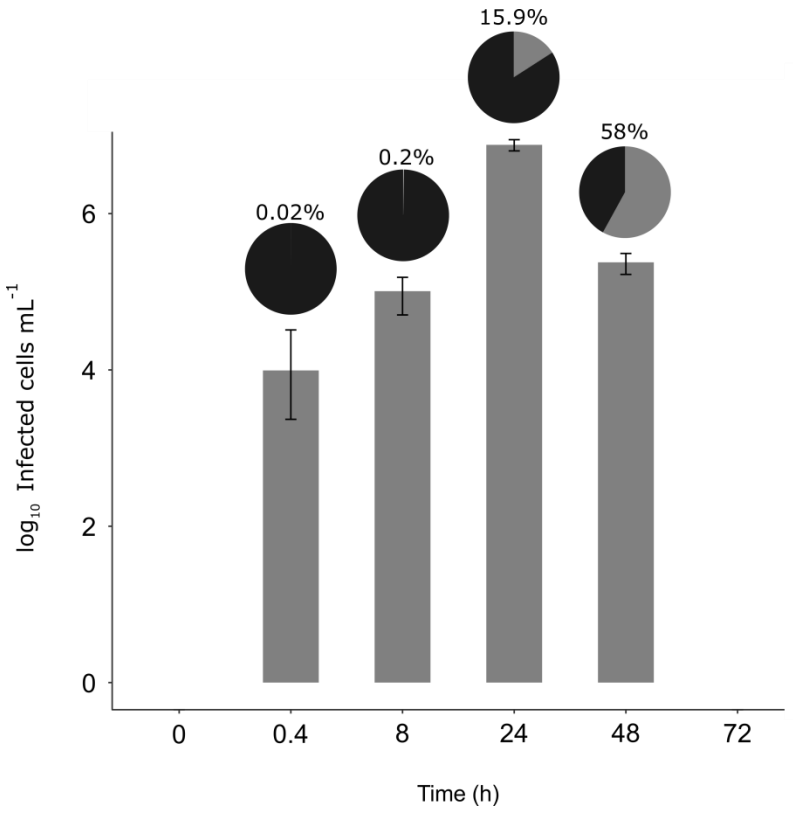


Figure 3. Dynamics of the infected cells. Bar plot shows the number of infected *O. tauri* at each time. Pie charts on top of each bar show the percentage of infected cells with respect to the total *O. tauri* abundance.

1.3.2. Dynamics and abundances of free OtV5 particles

We also used VirusFISH for the detection and quantification of free OtV5 particles produced during the infection and lysis of *O. tauri*. As mentioned above, the *O. tauri* culture is not axenic, so we performed a SYBRGold staining step to label all the dsDNA viruses present (green particles in Figure 4), which include OtV5, bacteriophages, vesicles and/or other artefacts of non-specific staining. Since a certain background can be observed in the micrographs, only the VirusFISH red signal that overlapped with a SYBRGold green fluorescence signal was considered a true OtV5 particle (yellowish particles in Fig. 4). Our results showed that before 24 h, at least 1 infection cycle had already

completed, as indicated by the slight, but detectable increase of OtV5 free particles and the slight decrease of *O. tauri* cells at 24 h. This is in agreement with the detection of 16% infected *O. tauri* cells at 24 h, which is higher than expected for the MOI used, as explained above. A drastic increase in viral abundance was observed after 24 h (Fig. 5 and Fig. S4), corresponding with the time the majority of cells were lysed. At 48 h the number of free viruses reached a plateau, likely because most viral production had already occurred. As expected, no increase in OtV5 particles was detected in the control flasks. The fraction of OtV5 within the total viral community labeled with SYBRGold ranged from 0.9% (±0.2%) at 0 h when viruses were inoculated to 72.1% (±5.6%) at 48 h when almost all *O. tauri* cells were lysed (Fig. 2 and Fig. 5).

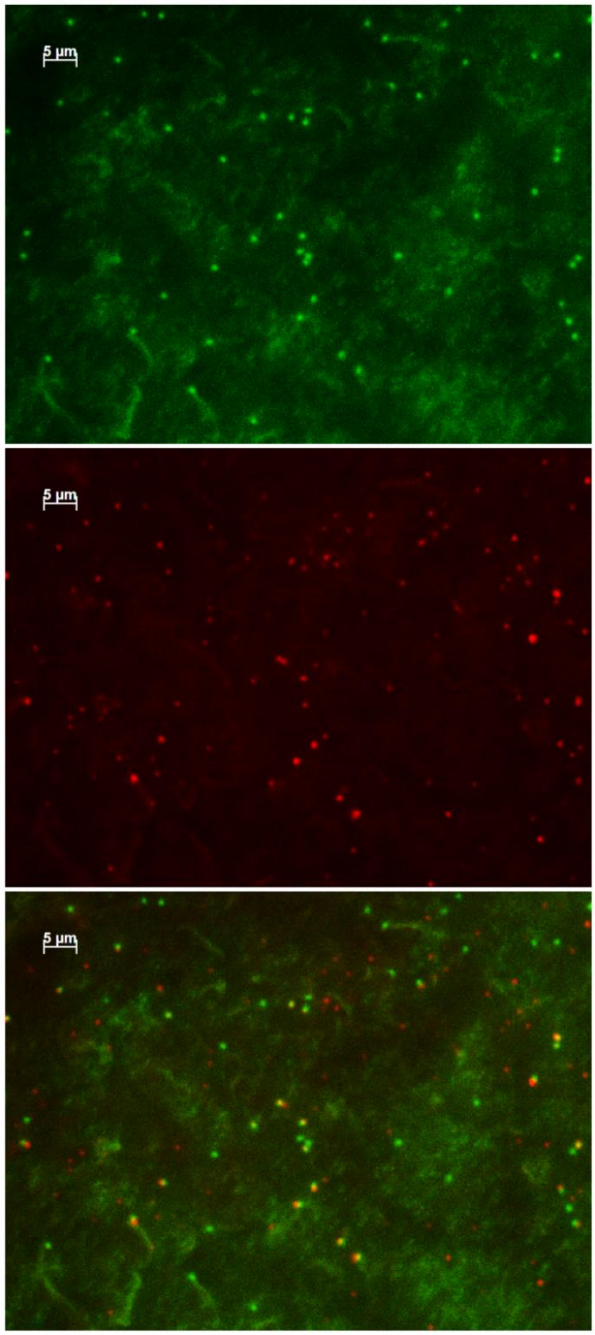


Figure 4. Micrographs of free viruses (at 48 h). Top: total viruses stained with SYBRGold. Center: VirusFISH labeled OtV5 viruses. Bottom: overlay of SYBRGold and VirusFISH signals for OtV5 viruses.

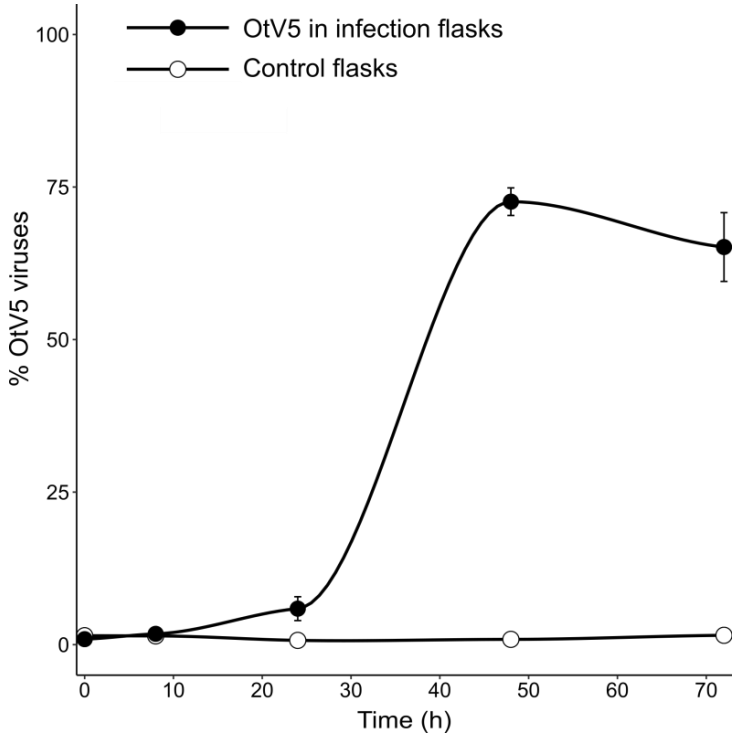


Figure 5. Dynamics of free viruses produced during the infection expressed as percentage of OtV5 with respect the total viral abundance. Counts were done by epifluorescence microscopy considering the overlay of signals.

1.3.3. Burst size

Applying the formula developed by Middelboe & Lyck (2002) that considers the increasing and decreasing abundances of free viruses and their hosts, respectively, we obtained a burst size value of 7±0.4 viruses cell^{-1}. To calculate this we considered that during lysis, the organic matrix released from the cells after 48 h trapped most of OtV5 particles on the filters (viral clouds). Therefore, to calculate the burst size we considered the number of free OtV5 particles at 24h ($1.9 \cdot 10^{6} \pm 3 \cdot 10^{4}$ viruses mL^{-1}) and, at 48 h, the sum of the number of free OtV5 ($2.5 \cdot 10^{7} \pm 3 \cdot 10^{6}$ viruses mL^{-1}) and the OtV5 particles within the viral clouds ($2 \cdot 10^{8}$ viruses mL^{-1}) to obtain a final value of viral abundance at that time.

Moreover, we used VirusFISH to corroborate the burst size value obtained by the classical method, as mentioned in the Material and Methods section. For this, we used the observed average area of a single free OtV5 particle (0.13 μm^2, n=30,000 viruses) from the Alexa594 fluorescence signal, and the average cellular area occupied by OtV5 virions at the maximum infection time-point before the major lysis occurred (24 h, 1.23 μm^2, n=90 areas). Using this approach we obtained a value of 9.5±0.3 viruses per each infected *O. tauri* cell, which is very close to the results obtained for the burst size.

1.4. DISCUSSION

Several studies have dealt with the virus–host relationships of the four clades of *Ostreococcus* spp. (*O. tauri*, *O. lucimarinus*, *O. mediterraneus* and clade B -Guillou et al., 2004-), and our knowledge on these systems is continuously expanding (Weynberg *et al.*, 2017). From these studies, only a few focused on the infection dynamics (e.g. Derelle *et al.*, 2008, 2017; Heath and Collins, 2016), and most of the work has been directed towards understanding the virus–host interaction at the molecular level (e.g. Derelle *et al.*, 2008; Weynberg *et al.*, 2011; Clerissi *et al.*, 2012), unveiling interesting information on the host resistance mechanisms to viruses (Thomas, 2011; Heath and Collins, 2016; Yau *et al.*, 2016). However, to understand the impact of viruses on the ecology of *Ostreococcus* spp. it is crucial to develop techniques that enable monitoring the host–virus interactions at the single cell level, with the ultimate goal to apply them in complex natural communities. We designed probes to detect OtV5, but the alignment of the probes with other Prasinovirus genomes showed that they can very likely label all 11 genome sequenced *Ostreococcus* spp. viruses (Table S2 and Table S3), except OtV6, which is evolutionarily distinct (Monier *et al.*, 2017). Thus, our technique may help in fostering our knowledge on the role of viruses in the control of the abundance of the cosmopolitan *Ostreococcus* spp.

Contrary to flow cytometry measurements and plaque-forming units assays, which only can give absolute cell and virus counts, VirusFISH allowed distinguishing and following the whole process of infection and shed light on what was happening previous to culture clearance, unveiling that infection was much more rapid than can be detected by cell or free virus counts. It showed that, despite most viruses seeming to have a period of latency after inoculation, some adsorbed producing a first discrete wave of infection after 8 h. At 24 h post-inoculation the infection percentage increased to a 16%, a quite low percentage if we consider that this process is followed by a surprising fast lysis of the culture only 24 hours later.

Another valuable application of VirusFISH was to determine the free viral particles released during infection, discriminating the true OtV5 from phages and other unspecific particles, improving the flow cytometry counts. Thus, we could estimate the burst size of an axenic culture (~7 viruses per cell). Also, the technique allowed corroborating the obtained burst size results by estimating the amount of viruses inside the cell at late stages of infection, giving similar results (~9.5 viruses per cell). If we compare these values with the 25 reported in Derelle *et al.* (2008) they do not extremely differ, despite there is a possibility that Derelle *et al.* (2008) could have overestimated the counts due to the fact that they used flow cytometry and could have counted phages as OtV5 particles. However, several studies have revealed that the experimental conditions affects the burst size value (Maat *et al.*, 2014; Maat and Brussaard, 2016), yielding a variation in the viral production among experiments. For instance, in *O. lucimarinus* (Zimmerman *et al.*, 2019) the infection of OlV7 virus differs depending on the growth light regimes. When the *O. lucimarinus* grows in optimal light conditions, the burst size is ~680 virus/host, but when the light conditions are suboptimal, and thus the cellular machinery is not working properly, the burst size decreases to ~50 virus/host. Therefore, burst size varies with the growing conditions. Nevertheless, burst size values may also vary

depending on the physiological state of the cells, and may decrease in the stationary phase of the culture (Demory *et al.*, 2017).

Furthermore, although it was not the goal of our study due to the tiny size of *Ostreococcus*, VirusFISH could be potentially used for visualizing the dynamics of the viruses within the eclipse phase in larger hosts (i.e. nanoeukaryotes), something that is not feasible with other methods like Transmission Electron Microscopy.

1.4.1. Methodological aspects to be considered for phototrophic eukaryotes and our particular O. tauri system.

One of the best fluorochromes to label gene probes is Alexa594 (Barrero-Canosa *et al.*, 2017), which emits red fluorescence when excited with orange light. However, the chloroplasts of photosynthetic microbes also emit red fluorescence under the same light, hampering the detection of viral signals. We solved this technical issue by removing the cellular pigments with a combination of alcohol treatments, as described in the materials and methods section.

The filter pore size also needs to be considered during VirusFISH experiments. *Ostreococcus* cells, although having a size of 1-3 µm, passed through a 0.6 µm filter, most likely because its cellular membranes are very flexible. This resulted in the loss of more than half of the cells during filtration. Consequently, we recommend the usage of filters with a pore size of 0.4 µm or 0.2 µm when working with picoeukaryotes. In our case, 0.2 µm pore size filters proved to be the best option, because, apart from completely retaining *O. tauri* cells, they allowed the visualization of viruses released from the lysed cells, and trapped in the organic matrix surrounding the cell debris (here referred as viral clouds) (Fig. 2, grey arrow). In contrast, these viral clouds could not be observed onto 0.4 µm filters, likely because the organic matrix passed through that pore.

1.4.2. Modifications of VirusFISH with respect to the published protocols of phageFISH and direct-geneFISH

VirusFISH represents a combination between phageFISH and the direct-geneFISH. It used CARD-FISH to identify the unicellular eukaryotic host, similar to phageFISH, and used a mixture of polynucleotide probes directly labeled with a fluorochrome to target viral genes, similar to the direct-geneFISH protocol. CARD-FISH was used because its signal amplification step enables the detection of cells with low ribosome content. Indeed, *O. tauri* and all Mamiellophyceae have a small cytoplasm due to the big size of the organelles (Yau *et al.*, 2016), and therefore their ribosomal abundance is low and CARD-FISH enhances the cellular visualization. We also incorporated a step of embedding the filters in agarose to avoid cell losses in downstream manipulations of the filter portions. Furthermore, because *O. tauri* lacks a cell wall, the permeabilization step was omitted. On the other hand, a treatment to completely remove cell pigments was required, as mentioned above. Finally, compared to the direct-geneFISH protocol, we reduced the Alexa594 fluorochrome volume to label the viral gene probes in order to reduce economical costs but obtaining equal optimal results (see details in the methods section).

1.4.3. VirusFISH vs other approaches to follow virus–host dynamics

Currently available methods to assess the dynamics between host and viruses during infection are i) the frequency of visibly infected cells (FVIC) (Wommack and Colwell, 2000), ii) Real Time PCR (RT-PCR) (Monier *et al.*, 2017) of viral genes and iii) the plaque assay, for counting plaque forming units (PFU) (Brussaard *et al.*, 2016). Compared with these methods, VirusFISH brings further advantages. For example, FVIC reports the fraction of infected host cells, but only detects those cells in the late stage of infection. PFU and RT-PCR describe the infection stages, but they lack the ability to measure the fraction of infected cells. With the exception of RT-PCR, which uses virus-specific primers, none of the three methods can identify the host or the viruses. In comparison, VirusFISH allows

the: i) identification of both host and virus, using 18S rRNA and viral genes specific probes, feature particularly advantageous in non-axenic cultures of unicellular eukaryotes or in environmental samples; ii) quantification of the total and relative abundance of the host cells; iii) quantification of the total and relative abundance of virus infected cells, independent of the stage of infection and; iv) quantification of released viral particles. Furthermore, VirusFISH can potentially be used to discriminate the different stages of infection, in a manner similar to phageFISH.

Some other approaches have arisen in the last decade to unveil virus–host interactions, like the polony method (Baran *et al.*, 2018) or the microfluidic digital PCR (Tadmor *et al.*, 2011). The novel polony method is a culture independent technique based on a single molecule PCR. Using degenerate primers it allows the determination of the abundance of a given viral group and its degree of diversity, discriminating between different viral families or genera, and their host. This high-throughput approach has enabled the quantitative assessment of thousands of viruses in a single sample from both aquatic and terrestrial environments (Baran *et al.*, 2018). Thus, the polony method is a powerful approach to detect virus–host interactions in a cost-effective and relatively simple manner, but similar to VirusFISH, it requires the knowledge of the hosts and the genome of the viral target to design the probes. However, although the VirusFISH approach is not as high-throughput as the polony method, it has the advantage that it allows monitoring and visualizing a particular viral infection, so we can see when the infection is taking place, how many cells are infected at different times and how the infection progresses.

Also, VirusFISH allows quantification of the number of viruses that are being produced during an infection event, and the estimation of burst size values. Likewise, it allows distinguishing temperate virus infections (single viral copy detection in a cell) from lytic infections (when it detects multi-copies in a cell). Moreover, since VirusFISH consists on microscopy observations it enables the

study of the heterogeneity of the infection within the host population, with the potential to extend its use to assessing those specific virus–host interactions in complex natural communities.

1.4.4. Free viruses abundance and estimates of burst size values

The abundance of free viruses has been traditionally assessed through i) plaque-assay count (Suttle and Chen, 1992), ii) transmission electron microscopy (TEM) of uranyl acetate stained virus particles (Malenovska, 2013)(Malenovska, 2013)(Malenovska, 2013) and; iii) by epifluorescence microscopy (Hennes *et al.*, 1995) or flow cytometry (Marie *et al.*, 1999) of SYBRGreen stained viruses. Each of the above methods have limitations: i) the plaque-assay is constrained to cultivable hosts and viruses; ii) TEM is a time consuming and expensive technique and; iii) SYBR staining followed by epifluorescence microscopy or flow cytometry does not distinguish between infective and non-infective viruses, making it impossible to identify the virus of interest within a complex viral community. With VirusFISH we achieved the detection of specific free viruses in a relatively fast way, with no requirements of specialized equipment, or extremely expensive reagents. Allers et *al.*, (2013) also applied phageFISH to visualize free viral particles, by immobilizing the viral lysate on glass slides, which can potentially lead to virus losses. With VirusFISH we tried to overcome this issue by collecting and counting the free viruses on 0.02 µm anodisc filters decreasing the risk of viral losses during the hybridization process due to a better retention. The proportion of OtV5 in relation to all viruses present in the non-axenic culture, was 72% at 48h (Fig. 5), indicating that bacteriophages represented a minor fraction of the viruses at that time point. Later the proportion of OtV5 decreased slightly, probably due to an increase in the proportion of bacteriophages. This proportion would have likely been higher if we had taken samples after 72h, since organic matter released by the lysed cells fueled bacterial growth (data not shown).

In summary, in this study we developed VirusFISH to detect the virus–host interaction of a picoeukaryotic system. This technique allowed us to visualize and follow the dynamics of the OtV5 viral infection of *Ostreococcus tauri* until the complete lysis of the culture. Also, VirusFISH enabled the calculation of the viral production during infection, discriminating OtV5 viruses from the phages present in the culture. Moreover, we demonstrated that VirusFISH can be used to calculate the burst size of hosts in non-axenic cultures. Also, our designed probes could potentially target most *Ostreococcus* viruses, except for OtV6, representing a valuable tool to address virus–host interactions in these cosmopolitan marine picoeukaryotes. We strongly believe that VirusFISH presents great prospects to address infection dynamics in nature, and it will foster our understanding on the impact of viruses in eukaryotic populations. Furthermore, this technique can be easily implemented to any other model system.

REFERENCES

Baran, N., Goldin, S., Maidanik, I., and Lindell, D. (2018) Quantification of diverse virus populations in the environment using the polony method. *Nat. Microbiol.* **3**: 62–72.

Barrero-Canosa, J., Moraru, C., Zeugner, L., Fuchs, B.M., and Amann, R. (2017) Direct-geneFISH: a simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. *Environ. Microbiol.* **19**: 70–82.

Brussaard, C.P.D., Baudoux, A.-C., and Rodríguez-Valera, F. (2016) Marine Viruses. In, *The Marine Microbiome*. Springer International Publishing, Cham, pp. 155–183.

Cabello, A.M., Latasa, M., Forn, I., Morán, X.A.G., and Massana, R. (2016) Vertical distribution of major photosynthetic picoeukaryotic groups in stratified marine waters. *Environ. Microbiol.* **18**: 1578–1590.

Chow, C.-E.T., Winget, D.M., White, R.A., Hallam, S.J., and Suttle, C.A. (2015) Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front. Microbiol.* **6**: 265.

Clerissi, C., Desdevises, Y., and Grimsley, N. (2012) Prasinoviruses of the marine green alga *Ostreococcus tauri* are mainly species specific. *J. Virol.* **86**: 4611–4619.

Demory, D., Arsenieff, L., Simon, N., Six, C., Rigaut-Jalabert, F., Marie, D., et al. (2017) Temperature is a key factor in *Micromonas*–virus interactions. *ISME J.* **11**: 601–612.

Derelle, E., Ferraz, C., Escande, M.-L., Eychenié, S., Cooke, R., Piganeau, G., et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS One* **3**: e2250.

Derelle, E., Yau, S., Moreau, H., and Grimsley, N.H. (2017) Prasinovirus attack of Ostreococcus is furtive by day but savage by night. *J. Virol.* **92**: JVI.01703-17.

Guillard, R.R.L. and Hargraves, P.E. (1993) Stichochrysis immobilis is a diatom, not a chrysophyte. *Phycologia* **32**: 234–236.

Guillou, L., Eikrem, W., Chrétiennot-Dinet, M.-J., Le Gall, F., Massana, R., Romari, K., et al. (2004) Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155**: 193–214.

Heath, S.E. and Collins, S. (2016) Mode of resistance to viral lysis affects host growth across multiple environments in the marine picoeukaryote *Ostreococcus tauri*. *Environ. Microbiol.* **18**: 4628–4639.

Henderson, G.P., Gan, L., and Jensen, G.J. (2007) 3-D ultrastructure of *O. tauri*: electron cryotomography of an entire eukaryotic cell. *PLoS One* **2**: e749.

Hennes, K.P., Suttle, C.A., and Chan, A.M. (1995) Fluorescently labeled virus probes show that natural virus populations can control the structure of marine microbial communities. *Appl. Environ. Microbiol.* **61**: 3623–3627.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* **9**: 2386–2399.

Maat, D. and Brussaard, C. (2016) Both phosphorus- and nitrogen limitation constrain viral proliferation in marine phytoplankton. *Aquat. Microb. Ecol.* **77**: 87–97.

Maat, D.S., Crawfurd, K.J., Timmermans, K.R., and Brussaard, C.P.D. (2014) Elevated CO2 and phosphate limitation favor *Micromonas pusilla* through stimulated growth and reduced viral impact. *Appl. Environ. Microbiol.* **80**: 3119–3127.

Malenovska, H. (2013) Virus quantitation by transmission electron microscopy, TCID50, and the role of timing virus harvesting: A case study of three animal viruses. *J. Virol. Methods* **191**: 136–140.

Marie, Brussaard, Thyrhaug, Bratbak, and Vaulot (1999) Enumeration of marine viruses in culture and natural samples by flow cytometry. *Appl. Environ. Microbiol.* **65**: 45–52.

Middelboe, M. and Lyck, P. (2002) Regeneration of dissolved organic matter by viral lysis in marine microbial communities. *Aquat. Microb. Ecol.* **27**: 187–194.

Monier, A., Chambouvet, A., Milner, D.S., Attah, V., Terrado, R., Lovejoy, C., et al. (2017) Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proc. Natl. Acad. Sci.* **114**: E7489–E7498.

Noble, R. and Fuhrman, J. (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. *Aquat. Microb. Ecol.* **14**: 113–118.

Not, F., Latasa, M., Marie, D., Cariou, T., Vaulot, D., and Simon, N. (2004) A single species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **70**: 4064–72.

Pavlekovic, M., Schmid, M.C., Schmider-Poignee, N., Spring, S., Pilhofer, M., Gaul, T., et al. (2009) Optimization of three FISH procedures for *in situ* detection of anaerobic ammonium oxidizing bacteria in biological wastewater treatment. *J. Microbiol. Methods* **78**: 119–126.

Pernice, M.C., Forn, I., Gomes, A., Lara, E., Alonso-Sáez, L., Arrieta, J.M., et al. (2015) Global abundance of planktonic heterotrophic protists in the deep ocean. *ISME J.* **9**: 782–792.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., et al. (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**: 676–682.

Suttle, C.A. and Chen, F. (1992) Mechanisms and rates of decay of marine viruses in seawater. *Appl. Environ. Microbiol.* **58**: 3721–9.

Tadmor, A.D., Ottesen, E.A., Leadbetter, J.R., and Phillips, R. (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**: 58–62.

Thomas, R. (2011) Action des phycoDNAvirus sur les populations phytoplanctoniques (Mamiellophyceae): étude de la résistance aux infections virales. *http://www.theses.fr*.

Weinbauer, M., Bettarel, Y., Cattaneo, R., Luef, B., Maier, C., Motegi, C., et al. (2009) Viral ecology of organic and inorganic particles in aquatic systems: avenues for further research. *Aquat. Microb. Ecol.* **57**: 321–341.

Weynberg, K., Allen, M., and Wilson, W. (2017) Marine prasinoviruses and their tiny plankton hosts: a review. *Viruses* **9**: 43.

Weynberg, K.D., Allen, M.J., Gilg, I.C., Scanlan, D.J., and Wilson, W.H. (2011) Genome sequence of *Ostreococcus tauri* virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. *J. Virol.* **85**: 4520–4529.

Wommack, K.E. and Colwell, R.R. (2000) Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**: 69–114.

Yau, S., Hemon, C., Derelle, E., Moreau, H., Piganeau, G., and Grimsley, N. (2016) A viral immunity chromosome in the marine picoeukaryote, Ostreococcus tauri. *PLOS Pathog.* **12**: e1005965.

Yau, S., Krasovec, M., Rombauts, S., Groussin, M., Benites, L.F., Vancaester, E., et al. (2019) Virus-host coexistence in phytoplankton through the genomic lens. *bioRxiv* 513622.

Zimmerman, A.E., Bachy, C., Ma, X., Roux, S., Jang, H. Bin, Sullivan, M.B., et al. (2019) Closely related viruses of the marine picoeukaryotic alga Ostreococcus lucimarinus exhibit different ecological strategies. *Environ. Microbiol.* **21**: 2148–2170.

SUPPLEMENTARY INFORMATION

Supplementary tables

Table S1. Position of the 11 probes in the OtV5 genome and the nucleotide sequence of the primers used to synthetize the viral probes.

Probe	Probe position in genome (bp) (Start / End)	Forward Primer sequence	Reverse Primer sequence
#1	102,846 / 103,145	CGAAGACGACAGCGAAGACCA	TCAGCCCAGGCGTTCTCTTGA
#2	103,148 / 103,447	GCCAAAAAAAGGGCGGCTGGGA	GCGAACTCATGACGGACTGTAC
#3	103,459 / 103,758	GCATCGTGGATGAGTTGCTCAA	TTGGTGGCGATGAGGCCAA
#4	103,765 / 104,064	TAAGGAGCCCCTCCCTGAT	GCGTTCTTTGTGCTAATTTCGC
#5	104,179 /104,478	GGGAAGCCAATTTGTTGGCGTG	AAACCTCGCGATGTTGGCTGCG
#6	104,479 /104,778	TGGGCACGACCCTTCTTCATCA	CTCAAGTGTCGCCGATGGTCAA
#7	104,792 / 105,091	CGATCACTTTTTCGGAGAGTTG	CATGATCCCCTTTGTTGGTTTA
#8	105,174 / 105,473	TATTTGGAATACCCTTACCCGT	TAAAGAGTGTTCGTTTCGTCAC
#9	105,631 / 105,930	GAGGAGGAAACGCTTTCCAG	CGCTACCTCCTGATCAAGAAGA
#10	105,941 / 106,240	TATGGCCCCTTCTCCGAGAAA	CGCGGAACTTTCTGAATTCCTC
#11	106,545 / 106,844	AGTTCTGCGGGGTGCGCAAA	ACGAGCTTGGTGAGGGCCTTA

Table S2. Alignment of our designed viral probes against Prasinovirus genomes and the outgroup PbCV1 (Chlorovirus) using BLAST. Shadow cells indicate identity higher than 80%, and coverage larger than 90%. Given than the probes are 300 bp long and all of them are combined for the hybridization, they may serve to visualize the virus-host interactions of all the *Ostreococcus* virus described in the table, except for OtV6. Accession numbers in Table S3.

	Probe 1		Probe 2		Probe 3		Probe 4		Probe 5		Probe 6		Probe 7		Probe 8		Probe 9		Probe 10		Probe 11	
	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)	Id (%)	Cov (%)
OtV5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
OtV1	97.7	100	98.7	100	98.7	100	99	100	98.3	100	99.3	100	99.7	100	98.3	100	98.7	100	98.7	100	96.7	100
OtV2	85.3	100	80.8	98	82	100	89.3	99	88.8	100	89.6	99	85.5	99	82.2	93	84.6	99	84.6	99	82	100
OtV6	77.4	100	0	0	74.8	90	83.2	99	78.6	93	82.2	99	75.3	97	0	0	0	0	0	0	84.1	35
OlV1	86.7	100	80.7	98	81.8	100	85.6	99	84	100	83	99	83	98	82	98	85.6	99	85.2	99	76.3	99
OlV2	85.3	100	84.2	98	82.3	100	91	99	89.1	100	89.6	99	85.5	99	83.6	93	99	84.6	84	99	80.7	100
OlV3	85.3	100	84.5	98	82	100	91.3	99	91.1	100	89	99	85.9	99	83.3	93	84	99	85	99	81	100
OlV4	0	0	79.7	87	81.8	100	85.6	99	85.3	100	82	99	82.7	98	82.3	98	85.6	99	86	99	74.4	100
OlV5	85	100	80.7	98	82.7	100	89.9	99	89.1	100	89.3	99	85.2	99	81.8	97	85	99	84.6	99	83	100
OlV6	85.3	100	80.6	97	82.7	100	89.9	99	89.4	100	89.3	99	85.2	99	81.8	93	85	99	84.6	99	82	100
OlV7	85.3	100	80.1	98	81.8	199	86	99	85.3	100	83.7	99	82.2	99	83.6	93	85	99	85.3	99	74.6	99
OmV1	95.7	100	98.3	100	98.3	99	99.7	99	97.7	100	99.7	100	99	100	99.3	100	99.3	100	99.3	100	98.3	100
OmV2	84	100	78	92	86.7	87	88.6	99	87.8	99	89	99	88.3	99	80.7	91	87	99	85.6	99	87.8	98
MpV1	74.8	99	81.6	25	69.3	90	73.7	76	71.8	93	76	82	81.5	59	0	0	86.5	35	0	0	76	25
BpV1	71.1	74	86.7	20	66.3	33	0	0	79.7	42	0	0	0	0	0	0	73.1	22	0	0	0	0
PbCV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Abbreviations: Id, Identity; Cov, Covery; OtV, *Ostreococcus tauri* virus; OlV, *Ostreococcus lucimarinus* virus; OmV, *Ostreococcus mediterraneus* virus; MpV, *Micromonas pusilla* virus; BpV, *Bathycoccus prasinos* virus; PbCV, *Paramecium bursaria* chlorella virus 1.

Table S3. Genbank accession of Prasinovirus genomes and the outgroup PbCV1 (Chlorovirus) used in Table S2.

Virus name	Accession number
OtV5	EU304328.2
OtV1	FN386611.1
OtV2	FN600414.1
OtV6	JN225873.1
OlV1	MK514405.1
OlV2	KP874736.1
OlV3	HQ633060.1
OlV4	JF974316.1
OlV5	HQ632827.1
OlV6	HQ633059.1
OlV7	MK514406.1
OmV1	KP874735.1
OmV2	Described in a preprint article (Yau et al., 2019) and obtained from the authors
MpV1	HM004429.1
BpV1	HM004432.1
PbCV	NC_000852.1

Abbreviations: OtV, *Ostreococcus tauri* virus; OlV, *Ostreococcus lucimarinus* virus; OmV, *Ostreococcus mediterraneus* virus; MpV, *Micromonas pusilla* virus; BpV, *Bathycoccus prasinos* virus; PbCV, *Paramecium bursaria* chlorella virus 1.

Supplementary figures

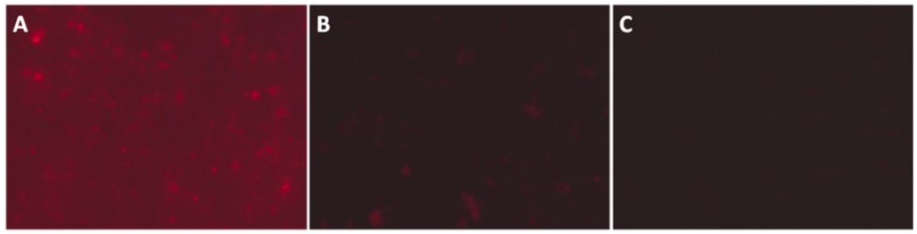


Figure S1. Cleaning test of *Ostreococcus tauri* 4221 culture for chlorophyll pigments removal. **A.** Negative control, no alcohol treatment. Chlorophyll is clearly visible. **B.** Cell culture treated for 1 hour with ethanol 97%. Chlorophyll is visually reduced, but not completely eliminated. **C.** Cell culture treated for 1 hour with ethanol 96% followed by 1 hour pure methanol. Chlorophyll is completely removed and almost no red background can be appreciated.

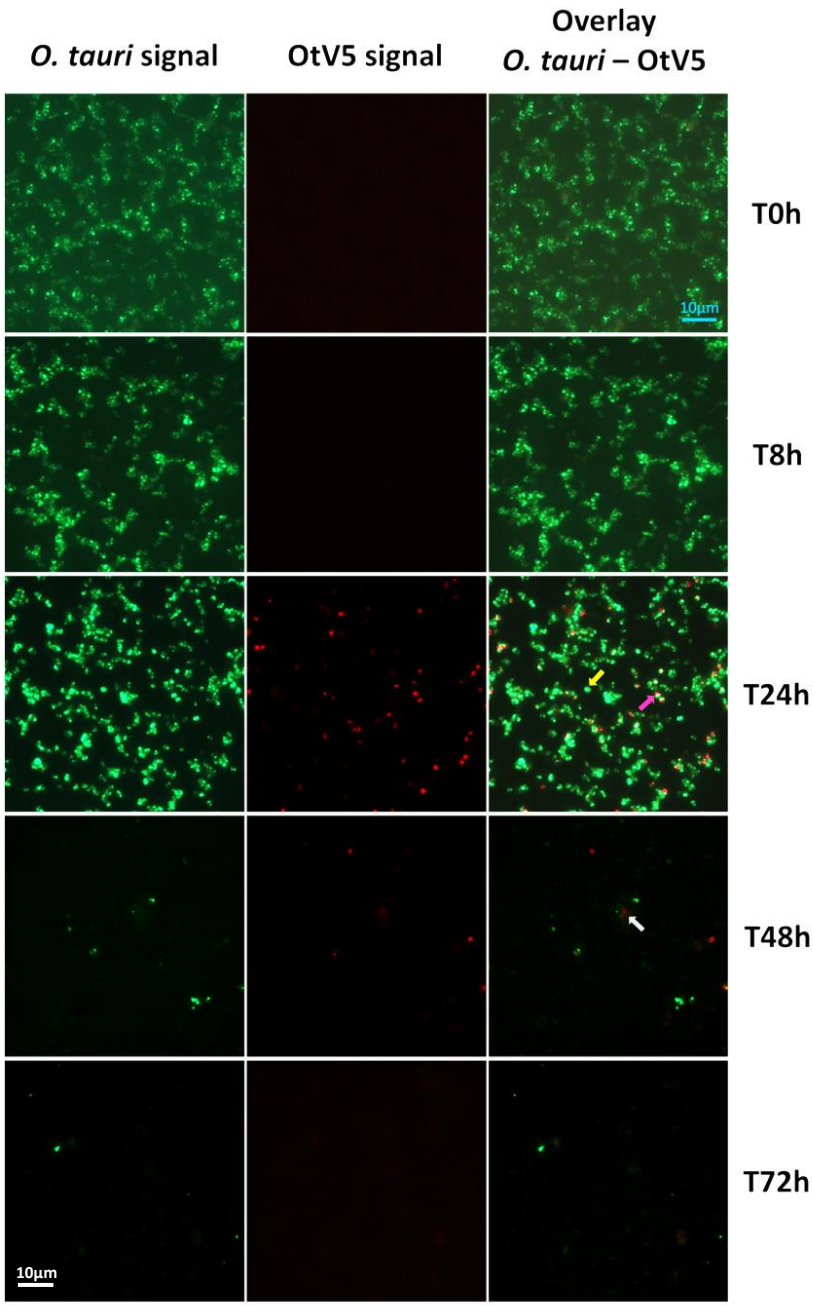


Figure S2. Micrographs of the evolution of the infection from time 0h to 72h. Left: *O. tauri* only. Centre: OtV5 only. Right column: overlay of *O. tauri* host cells in green (Alexa488) and virus in red (Alexa594). Yellow arrow: non-infected *O. tauri*; pink arrow: infected *O tauri*; grey arrow: cloud of viruses retained by the organic matter released during the lysis on the filter.

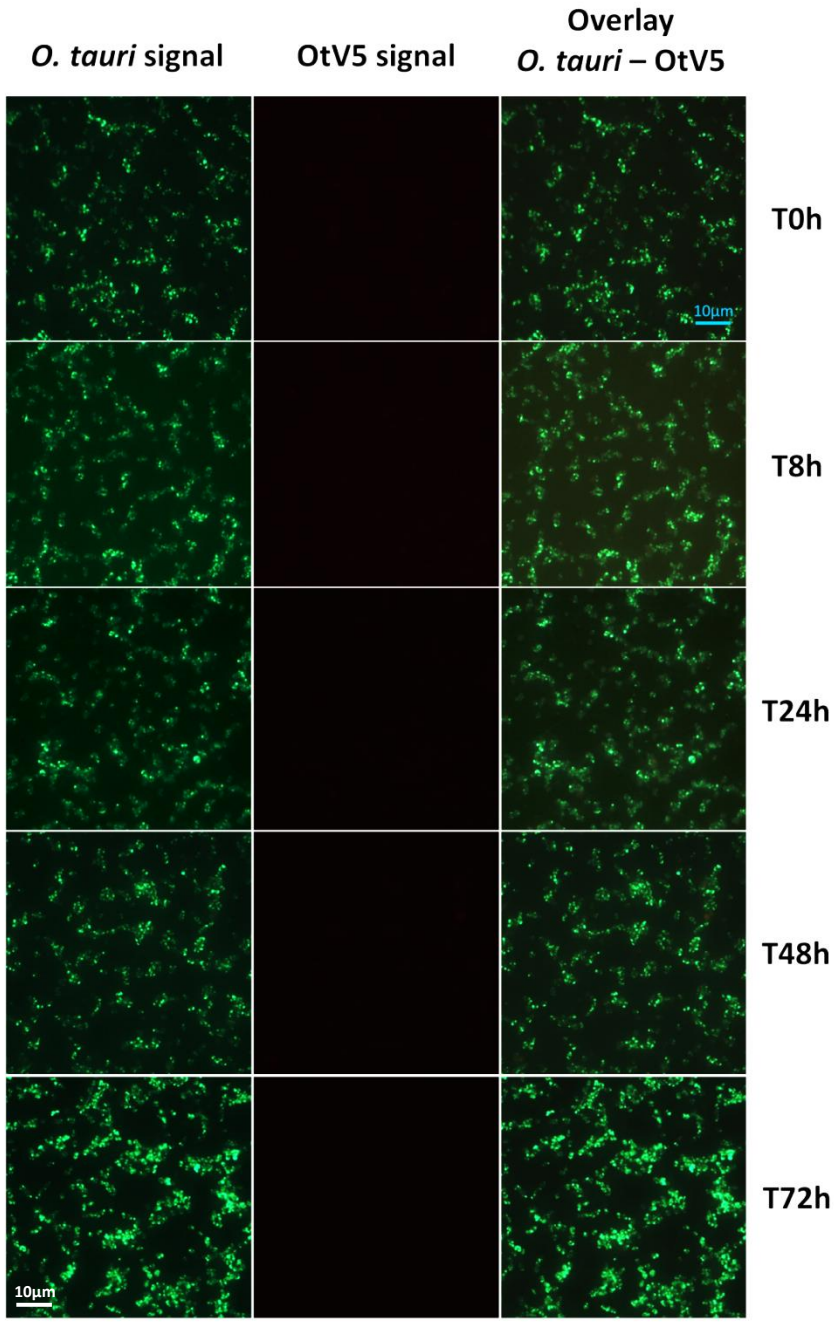


Figure S3. Micrographs of the evolution of the control flasks from time 0h to 72h. Left: CARD-FISH against *O. tauri* (Alexa488). Centre: VirusFISH against OtV5 viruses (Alexa594). Note no viral probes hybridation or false positives. Right column: overlay of both hybridization colors (Alexa488 and Alexa594).

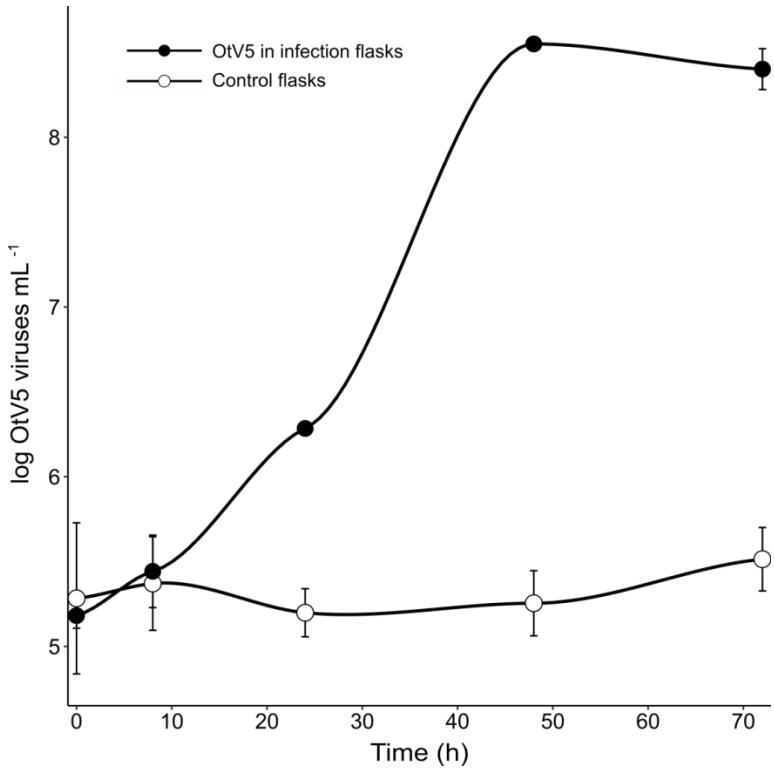
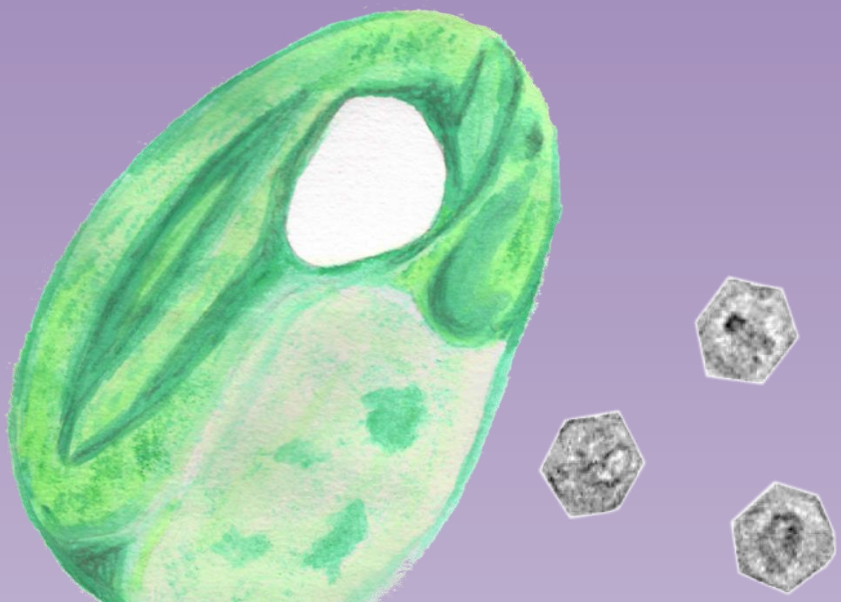
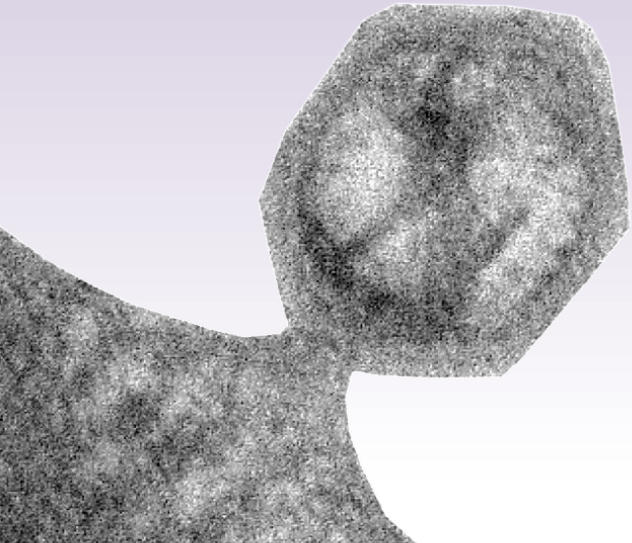


Figure S4. OtV5 particles abundance determined by VirusFISH on 0.02 µm anodisc filters. Black dots represent OtV5 particles in infection flasks. White dots represent the minimum basal false positive OtV5 viruses (mean of $3 \cdot 10^5$ virus mL^{-1}) detected in control flasks where no viruses were added.



CHAPTER 2

Seasonal dynamics of *Ostreococcus* spp. viral infection at the single cell level using VirusFISH



Castillo, Y.M., Sebastián, M., Forn, I., Yau, S., Morán, X.A.G., Alonso-Sáez, L., Arandia-Gorostidi, N. and Vaqué, D. Seasonal dynamics of *Ostreococcus* spp. viral infection at the single cell level using VirusFISH. Soon to be submitted.

ABSTRACT

Ostreococcus (Mamiellophyceae) is a cosmopolitan marine genus of phytoplankton found in mesotrophic and oligotrophic waters. With a picoplanktonic size (0.8-2.0µm), it is the world's smallest free-living eukaryote known to date, and it has been extensively studied as a model system to investigate viral–host dynamics in culture. Yet, the impact of viruses in naturally occurring populations of *Ostreococcus* is largely unknown. Here we used Virus Fluorescent *in situ* Hybridization (VirusFISH) to visualize and quantify the viral impact on this picoeukaryotic genus during a seasonal cycle in the central Cantabrian Sea (Southern Bay of Biscay), including coastal, continental shelf and offshore waters, both at surface and 50 m depth. The results showed that *Ostreococcus* was predominantly found during summer and autumn at both depths and all stations, representing up to 21% of the picoeukaryotic communities. Viral infection was only detected in surface waters, and its impact was variable but important from May to July and November to December, when up to half of the population was infected. Metatranscriptomic data available from the continental shelf station showed that the main active species in the samples was *Ostreococcus lucimarinus*. This work constitutes the proof of concept that VirusFISH can be used to quantify the impact of viruses on the population of key microbes among complex natural communities.

KEYWORDS

VirusFISH; natural communities; *Ostreococcus* spp.; viruses; infection dynamics.

2.1. INTRODUCTION

Half of the global primary production in our planet occurs in the sea, mostly by planktonic microorganisms that represent only a small fraction of the global primary producers biomass (Field, 1998). Picophytoplankton (cells <3 µm), that includes unicellular cyanobacteria and photosynthetic picoeukaryotes, are major contributors to phytoplankton biomass and primary production in marine systems (Massana, 2011). Unicellular cyanobacteria like *Prochlorococcus* and *Synechococcus* have been long assumed to be the main players due to their numerical dominance in oligotrophic waters. Yet, it is increasingly recognized that the low abundance photosynthetic picoeukaryotes dominate primary biomass and production due to the larger cell size and faster rates (Worden *et al.*, 2004).

Besides their role in primary productivity, photosynthetic picoeukaryotes are the prey for the dominant grazers in the ocean, i.e. larger nanoflagellates and ciliates (Fogg and Thake, 1987; Worden and Not, 2008), who act as link with higher levels of the trophic food web (Massana, 2011). The abundance of photosynthetic picoeukaryotes depends on their growth, which is subjected to their adaptation to environmental variables (bottom-up control), and on their mortality losses due to grazers and viruses (top-down control). This top-down control is considered to largely influence the population dynamics of different picoeukaryotes, but most studies have focused on the effect of viruses and grazers from a bulk community point of view (Bec *et al.*, 2005; Mojica *et al.*, 2016), without taking into account that virus–host relationships are rather specific (e.g. Lara et al., 2017; Sandaa & Larsen, 2006).

The detection of specific viruses infecting their host is not trivial, and it is one of the major challenges in viral ecology. Due to the absence of a universal phylogenetic marker for viruses, the assessment of virus–host interactions in natural systems has been obtained from metagenomics, which are boosting our knowledge on virus diversity and their potential hosts (e.g., Castillo *et al*., 2019, Mizuno *et al*. 2013, Roux *et al.*, 2017), and through PCR amplification of

conserved marker genes within specific viral families (Chen and Suttle, 1995; Larsen *et al.*, 2008; Lehahn *et al.*, 2014). Metatranscriptomic data has also been used to follow some infection dynamics (Zeigler Allen *et al.*, 2017). Yet, the impact of viruses on their host populations is hard to infer using those techniques.

Due to the mounting evidence on the role that viruses may play in bloom termination, quite a lot of attention has been recently paid to bloom forming species, like *Emiliania Huxleyi*. It has been shown that more than 60% of *E. huxleyi* cells may be infected at the demise phase of the bloom (Vardi *et al.*, 2012). However, under high host cell abundances, as in these blooms, the probability of encounter of a virus with its host is high, resulting in a fast viral propagation through the host population (Brussaard *et al.*, 2004; Baudoux, 2007). Nevertheless, the impact of viruses on hosts that may form occasional blooms (Zingone *et al.*, 1999; O'Kelly *et al.*, 2003; Countway and Caron, 2006; Johannessen *et al.*, 2017) but are generally present at low abundances has been little explored.

However, there are several studies focusing on the occasional blooming species *Micromonas pusilla,* a member of the Mamiellaceae family. For example, Cottrell *et al.* (1991) looked at the abundance of *M. pusilla* viruses (MpV) at different locations in October 1990 using the most-probable number approach (MPN) on cultured hosts. Cottrell *et al.* (1995) went a step further to infer temporal dynamics by doing a weekly sampling from January to April 1993 in different sampling sites, using also the MPN approach. Later, Zingone *et al.* (1999) reported the occurrence and temporal patterns of viruses infecting specific marine phytoplankton cells in relation to the abundance of their host, combining the MPN and epifluorescence microscopy approaches to calculate the virus and host abundances over three years. Yet, in these studies the infection rates are inferred from the viral abundances obtained through MPN using cultured hosts and, therefore, they are questionable due to the intraspecific diversity of natural algal and viral populations. Johannessen *et al.* (2017) also looked at the

dynamics of haptophytes and their viruses over time, combining flow cytometry to calculate the abundances and pyrosequencing to reveal both the diversity of haptophyte OTUs and algal viruses in seawater samples. Nevertheless, these approaches do not provide any information about the interaction between viruses and hosts.

Recently, a promising method to detect, visualize and follow viral–host dynamics, Virus Fluorescent *in situ* Hybridization (VirusFISH), was implemented to monitor viral infection on *Ostreococcus* (Castillo *et al.*, Chapter 1) using the model system *Ostreococcus tauri* – OtV5 virus. The genus *Ostreococcus* (Mamiellaceae) comprises cosmopolitan marine phytoplankton taxa which can be ubiquitously found from the coast to the open ocean, and from mesotrophic to oligotrophic waters (Derelle *et al.*, 2006). VirusFISH combines a Catalyzed Reporter Deposition Fluorescent *in situ* Hybridization detection of different *Ostreococcus* species with the general OSTREO01 probe (Not *et al.*, 2004) and viral probes originally designed for the detection of *Ostreococcus tauri* virus OtV5, but which targets also several virus that infect different species of *Ostreococcus* (Chapter 1, Table S3).

The purpose of this study is to validate the VirusFISH technique to assess the impact of Prasinoviruses on natural populations of *Ostreococcus* over a seasonal cycle. To this end, we took monthly samples at the Bay of Biscay (Cantabrian Sea), from coastal, continental shelf and offshore waters. With VirusFISH we are able to detect and quantify the proportion of natural infected *Ostreococcus* at any given time. The availability of metatranscriptomic data from the surface continental shelf station, allowed us to identify the viruses that were actively infecting the different *Ostreococcus* spp. With this study we demonstrate that VirusFISH is a powerful tool for studying virus–host interactions in the environment, which is crucial to advance in our understanding on the role of viruses in controlling their hosts abundances, and thus their impact in the ecology of marine microbes.

2.2. MATERIALS AND METHODS

2.2.1. The study area

Seawater samples for VirusFISH were taken from the Cantabrian Sea (Southern Bay of Biscay, Gijón, Spain) monthly between January and December 2012. Samples were collected at surface and at 50 m depths from 3 different stations: coastal (E1; 20m maximum depth (max depth); 43.58° N, 5.61° W), continental shelf (E2; 100m max depth; 43.67° N, 5.58° W) and offshore (E3; 150m max depth; 43.78° N, 5.55° W) (Fig. 1).

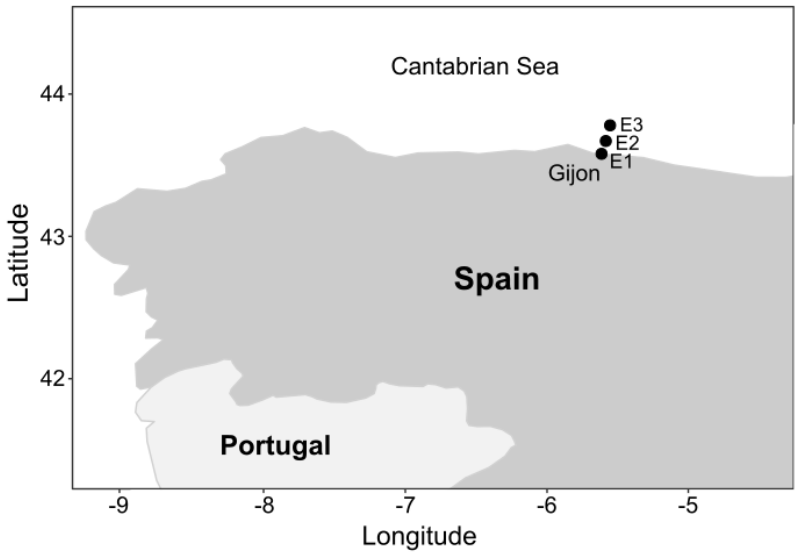


Figure 1. Location of the sampling stations. Abbreviations: E1: station 1, coastal; E2: station 2, continental shelf; E3: station 3, offshore. Samples were taken at 0 and 50m depth, except for E1 where the maximum depth was 20m.

2.2.2. The target system

Ostreococcus is the world's smallest free-living eukaryote known to date (0.8-2.0µm) (Derelle *et al.*, 2006). It is categorized in four different strains (from A to D) that correspond to the species *O. lucimarinus*, clade B (no species name assigned yet), *O. tauri* and *O. mediterraneus*, respectively, and three ecotypes (coastal, oceanic, and deep water). *O. tauri* and *O. mediterraneus* are generally

restricted to the surface layer of the ocean (0-5 m), while *O. lucimarinus* extends from surface to 65 m depth (Rodríguez *et al.*, 2005), and all three come from high-light, nutrient rich, cooler and more coastal waters (Demir-Hilton *et al.*, 2011). In contrast, Clade B is a deep and low-light adapted species isolated from the bottom of the euphotic zone (90-120 m) (Rodríguez *et al.*, 2005; Demir-Hilton *et al.*, 2011). The number of sequenced *Ostreococcus* viruses is constantly increasing, and nowadays the complete genomes of many Prasinoviruses are available (e.g Derelle *et al.*, 2008, 2015; Weynberg *et al.*, 2011; Monier *et al.*, 2017). Yet, nothing is known about these virus–host relationships *in situ*, and how they change over a temporal scale.

2.2.3. Environmental and biological setting

Seawater samples for Chlorophyll *a*, inorganic nutrients and picoeukayotic abundance were collected monthly between January and December in 2011 and 2012. Samples for VirusFISH were collected monthly in 2012. Temperature and salinity were measured by a SeaBird 25 CTD. Samples for Chlorophyll *a* (Chla) concentrations were collected by filtering 200 mL subsamples onto 0.2 µm polycarbonate filters. Samples for Chla concentration in the picosize fraction (pChl, <2µm) were obtained by sequentially filtering 200 ml of seawater through 2 µm and 0.2 µm filters. Filters were kept frozen at −20°C and processed within two weeks, as explained in Calvo-Díaz and Morán (2006). Chla was later determined in the laboratory by fluorimetry (Arandia-Gorostidi *et al.*, 2017). Picoeukaryotic abundances were acquired by flow cytometry using 1.8 mL subsamples fixed with glutaraldehyde (1% final concentration) as described in Morán *et al*. (2018).

2.2.4. VirusFISH: sample preparation, labeling and analysis

Samples for VirusFISH were collected only from January to December 2012. Seawater was pre-filtered through 0.8 µm pore-size cartridges (PALL Corporation, East Hills, NY, USA) to remove predators. After, 4 mL samples were

fixed with 3% formaldehyde freshly pre-filtered and cells collected onto 0.2 µm filters (Arandia-Gorostidi *et al.*, 2017). Filters were kept at -80°C until their analysis. Cells and viruses were hybridized and analyzed as described in Materials and Methods in Chapter 1. Briefly, samples were treated with alcohols to remove pigments, then cells were hybridized with the OSTREO01 probe for CARD-FISH, labeled with Alexa488, and after, viruses were hybridized with the 11 viral probes designed for *Ostreococcus* viruses labeled with Alexa594 (see Chapter 1, materials and methods, Table S2 for more details). Images were manually acquired using a Zeiss Axio Imager Z2m epifluorescence microscope (Carl Zeiss, Germany) connected to a Zeiss camera (AxioCamHR, Carl Zeiss MicroImaging, S.L., Barcelona, Spain) at x1000 magnification through the AxioVision 4.8 software. *Ostreococcus* cells were observed by epifluorescence microscopy under blue light (475/30 nm excitation, 527/54 BP emission, and FT 495 beam splitter) and *Ostreococcus* viruses under orange light (585/35 nm excitation, 615 LP emission, and FT 570 beam splitter). All pictures were taken using the same intensities and exposure times (300 ms for the blue light and 1 s for the orange light). For each sample, 4 random transects, between 6 and 10 mm each, were performed to analyze and count.

2.2.5. Metatranscriptomics: sample preparation and processing

Eight samples for metatranscriptomics were collected from the continental shelf at the surface in April, May, July and November, in 2011 and 2012, as reported in (Alonso-Sáez *et al.*, 2018). Briefly, samples were collected and immediately pre-filtered using 3 µm polycarbonate filters and cells retained onto 0.22 µm polycarbonate filters (Millipore). 0.22 µm filters were placed with 2 mL of RLT buffer (Qiagen) and 10 µl beta-mercaptoethanol in Whirl-Pak bags, flash frozen in liquid nitrogen, and kept at -80°C until analysis. The processing of the RNA was carried out as described in (Alonso-Sáez *et al.*, 2018). Sequencing depth is shown in Table S1 and the analyzed data corresponds to the non-rRNA reads.

2.2.6. Identification of *Ostreococcus* spp. and *Ostreococcus* virus sequences in metatranscriptomes

Metatranscriptomic reads, previously quality trimmed and cleaned of rRNA sequences (Alonso-Sáez *et al.*, 2018), were screened for *Ostreococcus* spp. (OS) and *Ostreococcus* virus (OV) sequences. First, a BLASTn database was constructed of the four *Ostreococcus* species nuclear genomes (*O. tauri* RCC4221, *O. lucimarinus* CCE9901, *Osterococcus* sp. RCC809 and *O. mediterraneus* RCC2590) and the 13 complete *Ostreococcus* spp. virus genomes sequenced to date. The Genbank accession numbers of the genomes used were as follows. *O. tauri*: CAID01000001.2–CAID01000020.2, *O. lucimarinus*: CP000581.1–CP000601.1, OtV1: FN386611.1, OtV2: FN600414.1, OtV5: EU304328.2, OtV6: JN225873.1, OlV1: MK514405.1, OlV2: KP874736.1, OlV3: HQ633060.1, OlV4: JF974316.1, OlV5: HQ632827.1, OlV6: HQ633059.1, OlV7: MK514406.1 and OmV1: KP874735.1. The *Ostreococcus* sp. RCC809 genome was obtained from the JGI Genome portal (https://genome.jgi.doe.gov/portal/ – accessed 28 February 2014). *O. mediterraneus* and OmV2 genomes are described in a preprint article (Yau *et al.*, 2019) and were obtained from the authors. Second, the metatranscriptomic reads were queried against OS and OV genomes by BLASTn (BLAST 2.2.26+), accepting high scoring pairs with e-value <1e−5, identity >75% and query coverage >75%. This nucleotide identity cut-off was chosen as it corresponds to the average nucleotide identity between *Ostreococcus* spp. (*O. tauri* and *O. lucimarinus*), as well as between representatives of *Ostreococcus* virus clades (OtV5 and OtV6), and thereby avoids retrieving reads that originate from related Mamiellophyceae and prasinoviruses. Average nucleotide identities were calculated with the ANI server (http://enve-omics.ce.gatech.edu/ani). Third, metatranscriptomic reads matching OS and OV genomes from each sample were counted, assigned to the species corresponding to the top BLASTn hit. Finally, *Ostreococcus* spp. and *Ostreococcus* virus read counts were expressed as counts per 100 000 reads to adjust for variation in per sample sequencing depth.

2.2.7. Transcriptome coverage of *Ostreococcus* viruses

To determine which regions of the viral genomes were expressed, all metatranscriptomic reads were aligned to the reference genome of the model *Ostreococcus* virus strain (OtV5 virus) using BWA version 7.17 (Li and Durbin, 2009) with default parameters. The resulting alignment was visualized in IGV version 2.5.3 (Robinson *et al.*, 2011).

2.2.8. Statistical analysis

Correlation analyses were performed using Pearson correlation. Statistical analyses were performed with the JMP 9.0.1 (JMP®, Version 9.0.1. SAS Institute Inc., Cary, NC, 1989-2019.) or R 3.5.3 (R Development Core Team, 2016) softwares.

2.3. RESULTS

2.3.1. Characterization of the sampling site

During 2011 and 2012, seawater temperature ranged from ~12 °C in winter to ~21 °C in summer in surface waters, whereas salinity was rather constant throughout the year at an average of 35.7 PSU, with a small decline in 2012 in April in all the three stations (~35 PSU, Fig. S1). Chla concentration at the surface in 2011 and 2012 peaked during spring and autumn reaching values of ~1 µg L^{-1} for the three stations, and also in summer for E1. At 50 m depth, temperature and salinity were quite stable throughout both years. In contrast, Chla showed a peak in June at E2 in 2011, and in May at E3 in 2012, and a peak in late summer for both stations in 2012 (Fig. S1). Between June and November there was a deep chlorophyll maximum (DCM) around 40-50 m in the continental shelf and offshore stations. As expected, Nitrate (NO_3) and phosphate (PO_4) concentrations were in general lower at surface than at the DCM for all stations and reached their maximum values during winter both in surface and 50 m

depth waters (Table S2). Moreover, the NO_3/PO_4 ratio had an average value of 10 (Table S2), which may indicate limitation by nitrate (Redfield *et al.*, 1963).

Abundance of picoeukaryotes (PE) in surface waters was in general two-fold higher than at 50 m depth (Fig. 2). At the surface, PE reached maximum abundances in April and November for all the three stations, with E1 also having high values in summer, coincident with the peak in Chla. At 50 m depth, PE were almost absent during winter but from late spring to autumn oscillated between 5000 and 20,000 cells mL^{-1} (Fig. 2).

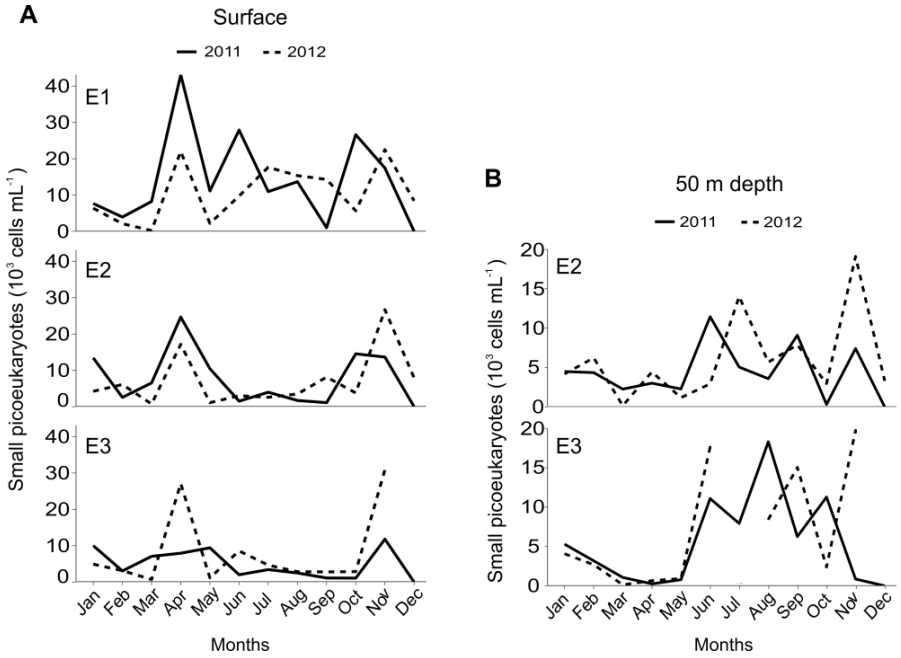


Figure 2. Small picoeukaryote abundances for coastal (E1), continental shelf (E2) and offshore (E3) waters during a two years period (2011 and 2012) at **A** the surface and **B** 50m depth. Note the difference in the y-axis between figures A and B.

2.3.2. Dynamics of *Ostreococcus* and its viral infection during an annual cycle

Using CARD-FISH and VirusFISH we followed the abundance of *Ostreococcus* cells and their viral infection during 2012. Infected *Ostreococcus* cells were visually detected from the red fluorescence of the VirusFISH labeled viruses (see methods section) overlapping with the green signal of the CARD-FISH *Ostreococcus* probe (Fig. 3).

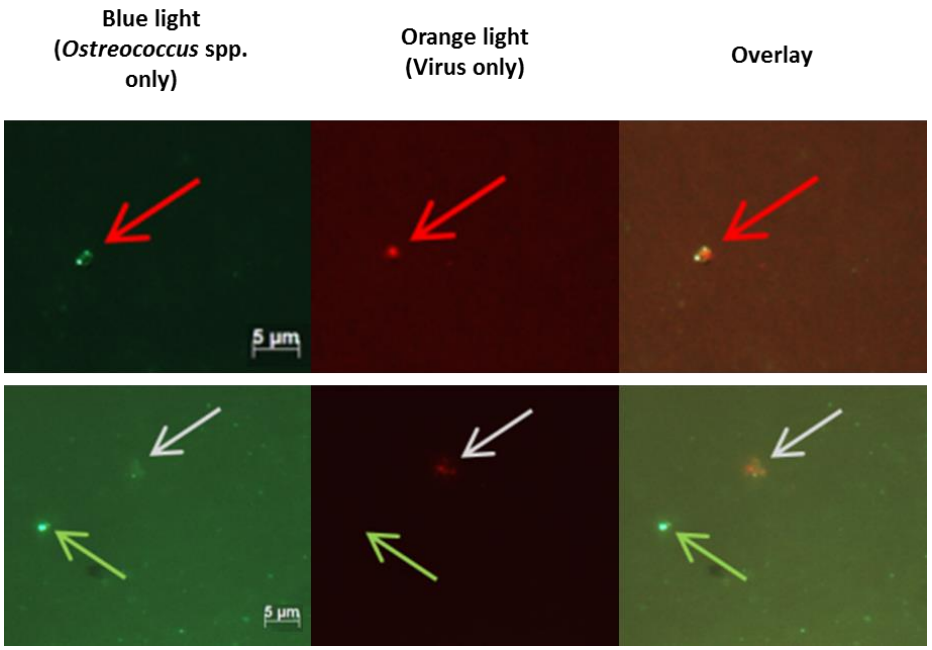


Figure 3. Micrographs of *Ostreococcus* cells in natural samples from the Cantabrian Sea. Upper panel: infected *Ostreococcus* cells (red arrows), in which the red signal of the VirusFISH labeled viruses can be easily seen. Lower panel: a healthy non-infected *Ostreococcus* cell (green arrows) and a completely lysed *Ostreococcus* cell releasing the viruses (gray arrows).

The contribution of *Ostreococcus* cells to the picoeukaryotic assemblages over the seasonal cycle ranged from 0 to 20.8% in surface waters, averaging 2.6%±0.73%, and from 0 to 8.9%, averaging 1.7%±0.45%, at 50 m depth (Table S3).

At surface waters in the coastal station (E1), *Ostreococcus* abundances started to increase in late spring and reached the highest values in summer (Fig. 4). At the continental shelf (E2) *Ostreococcus* cells displayed two relative maxima in July and November-December (Fig. 4), and at the offshore station (E3) we obtained similar results as in E2, with two relative maxima in July and November. Remarkably, *Ostreococcus* cells could not be detected in August at the two stations more distant from shore (E2 and E3), whereas they showed maximal abundances in the coastal station E1 (Fig. 4A). At 50m depth, *Ostreococcus* cells were also mainly found in summer and autumn, with the exception of October (Fig. 4B).

Although *Ostreococcus* abundances reached higher values in surface waters than at 50m depth, year-round average values were similar for both depths and among stations (i.e. E1 surface: 208.3±105 cells mL^{-1}; E2 surface: 127.9±53.6 cells mL^{-1}; E2 50m: 151.1±38.3 cells mL^{-1}; E3 surface: 121±54.1 cells mL^{-1}; E3 50m: 133.7±54.8 cells mL^{-1}).

In surface waters of E1, viral infection was observed in June, July, September, November and December, representing from 11 to 60% of the cells. In E2, the infected cells were visualized in late spring-early summer, representing from 7 to 50% of the cells. In E3, we could only detect infected cells in November, which accounted for 25% of the *Ostreococcus* population. Thus, the impact of viruses on *Ostreococcus* cells in surface waters of the Cantabrian Sea was variable, but infection took place mostly from May to June and from November to December (Fig. 4B, Table 1). Contrary to surface samples, at 50m depth no infected cells could be detected at any time (Fig. 4B).

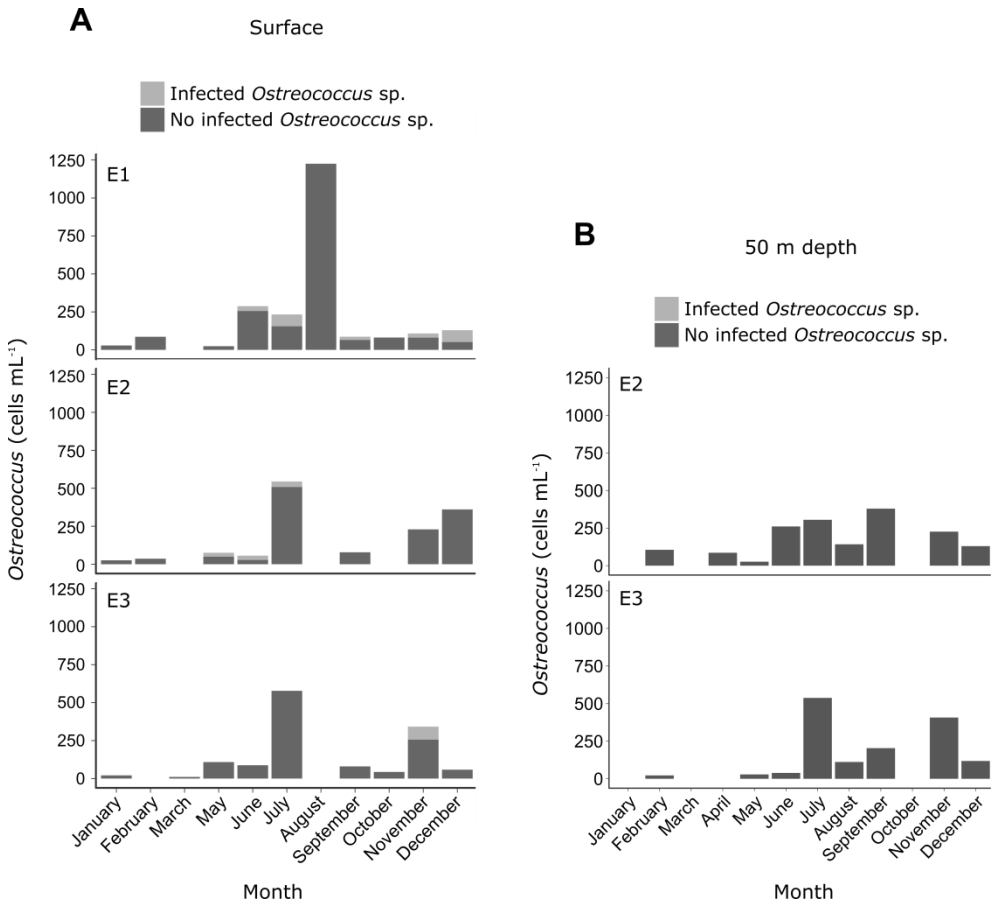


Figure 4. VirusFISH results for *Ostreococcus* cells abundance and infection by Prasinoviruses in 2012. **A.** Surface, **B.** 50m depth, in coastal (E1), continental shelf (E2) and offshore (E3) waters. Note: April data was not available for surface samples.

Table 1. Impact of Prasinoviruses (% of infected cells) on *Ostreococcus* populations

	E1	E2	E3
January	ND	ND	ND
February	ND	ND	–
March	–	–	ND
April	–	–	–
May	ND	33	ND
June	11	50	ND
July	33	7	ND
August	ND	–	–
September	25	ND	ND
October	ND	–	ND
November	25	ND	25
December	60	ND	ND

Abbreviatons: E1, Station 1; E2, station 2; E3, station 3; ND, non-detected, "–", no data.

There was a significant positive relationship (Pearson correlation analyses) between the abundances of *Ostreococcus* and both the number of infected cells and picoeukaryotes abundance (n=54, R=0.42, p-value=0.0016 and n=54, R=0.43, p-value=0.0013, respectively). Also, *Ostreococcus* abundance significantly increased with temperature (n=57, R=0.29, p-value=0.027) and the number of infected cells was inversely correlated with salinity (n=57, R=-0.32, p-value=0.016) (Table S4).

2.3.3. Detection of *Ostreococcus* and *Ostreococcus* virus in the metatranscriptomes

Both *Ostreococcus* (OS) and their viruses (OV) were detected in metatranscriptomic samples collected during 2011 and 2012 at the continental shelf station (E2), except for May and July 2011, when OV were not detected, coincident with very low abundances of host transcripts (Fig. 5).

The relative abundance of OS transcripts displayed a maximum in November, was second highest in April and was low in the spring and summer months of

May and July both in 2011 and 2012 (Fig. 5 lower panel). The relative abundance of OV transcripts followed the abundance of transcripts of their hosts, but depicting much lower abundances than their hosts. OV transcription was highest in April and November, and lowest in May and July. Also, OV transcript abundances were notably higher in 2011 than in 2012. Furthermore, the relative abundance of viral transcripts in relation to the abundance of host transcripts was higher in April 2011, pointing to a larger infection event at this sampling time point (Fig. 5).

Regarding the phylogenetic affiliation of the OS and OV transcripts, we found that the *Ostreococcus* assemblage maintained the same rank species abundance profile in all samples with *O. lucimarinus* as the most transcriptionally active species (51–91% of *Ostreococcus* reads), followed by *O. tauri* (6–47% of reads), while *Ostreococcus* sp. RCC809 and *O. mediterraneus* were minor contributors (1–4% of reads). This pattern was also reflected in the OV transcripts pool, with the dominance of *O. lucimarinus* viruses transcripts, followed by *O. tauri* viruses. *O. mediterraneus* viruses represented a minor fraction of the transcripts, whereas the only known virus infecting *Ostreococcus* sp. RCC809, OtV2 (Weynberg *et al.*, 2011), was not detected (Fig. 5).

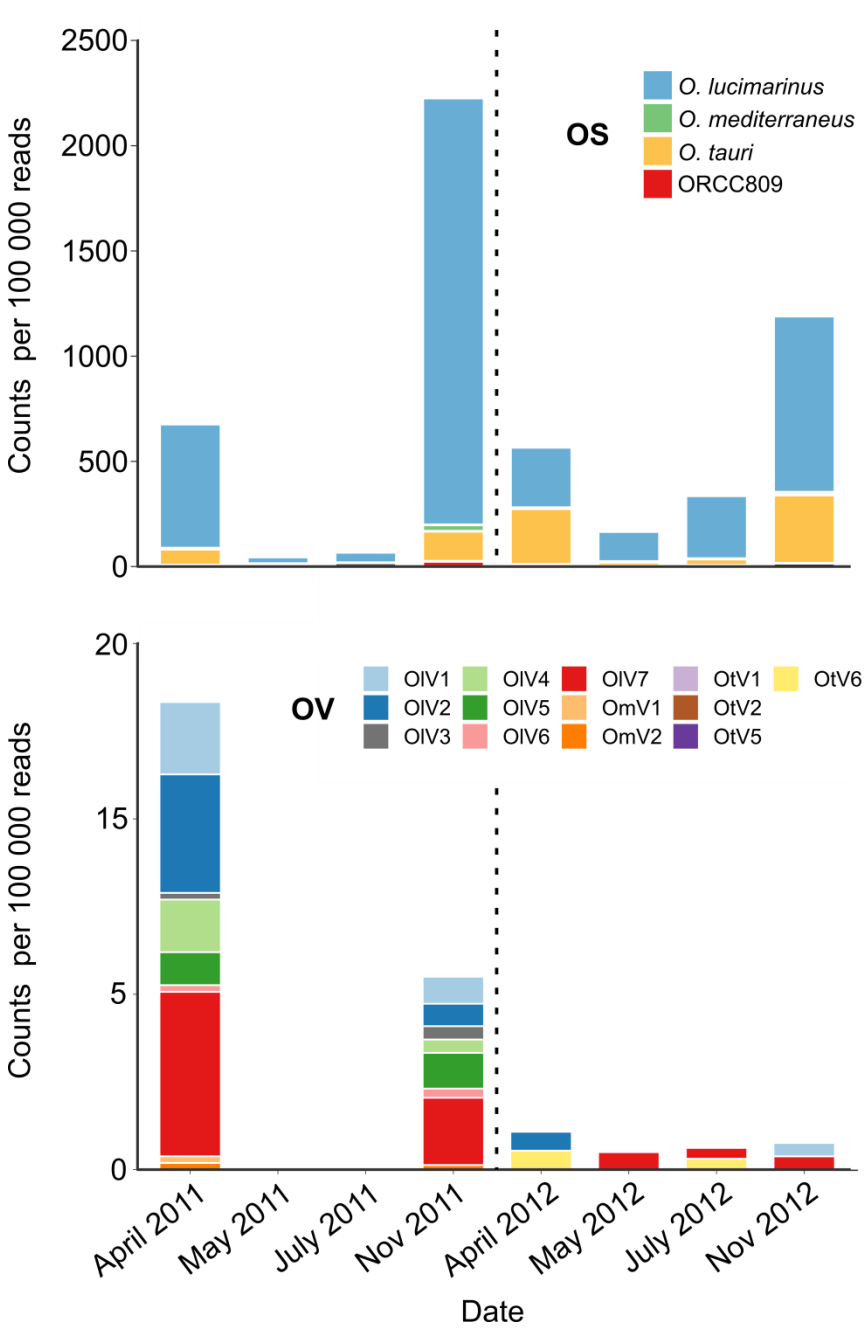


Figure 5. Relative abundances of *Ostreococcus* spp. (OS) transcripts (upper plot) and *Ostreococcus* viruses (OV) transcripts (lower plot) detected in metatranscriptomes. Note the difference in y-axis between the graphs. Abbreviation: Nov, November.

A comparison between the ratio of OV/OS transcripts versus the percentage of infected cells detected by VirusFISH showed consistent results, with a higher representation of viral transcripts in relation to the hosts in those samples where the number of infected cells was higher (Fig. 6).

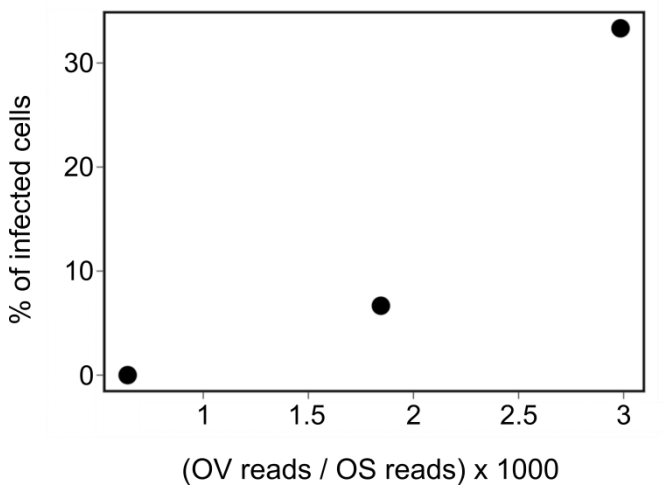


Figure 6. Relationship between the ratio of *Ostreococcus* virus transcripts in relation to *Ostreococcus* transcripts (viral transcriptional activity) and the percentage of infected *Ostreococcus* spp. obtained with VirusFISH. Abbreviations: OS, *Ostreoccocus* cells; OV, *Ostreococcus* virus.

2.3.4. Transcriptome coverage of *Ostreococcus* viruses

To determine which genomic regions of the *Ostreococcus* virus were being transcribed in the samples, all metatranscriptomic reads were aligned to the model *Ostreococcus* virus strain, OtV5, which is the virus that has received the most extensive molecular characterization (Derelle *et al.*, 2008; Yau *et al.*, 2016; Derelle *et al.*, 2017). This showed low read coverage (maximum 10 reads) along the genome length, but found DNA replication and capsid assembly genes. The most highly expressed gene was the major capsid protein, which is involved in the viral capsid assembly and likely reflect late stages of infection (Fig. S2).

2.4. DISCUSSION

Most studies dealing with the impact of viruses on microbial populations have been carried out using a bulk community approach (e.g. Mojica *et al.*, 2016), without taking into account the specificity of viral–host interactions. A few exceptions are studies that focused on the dynamics and abundances of specific viruses and hosts in nature, using the MPN approach and cultured hosts to quantify the amount of infective viruses (Cottrell and Suttle, 1991, 1995; Zingone *et al.*, 1999). These studies also estimated the proportion of hosts lysed per day (Cottrell and Suttle, 1995) or the levels of infection (Zingone *et al.*, 1999), but the numbers were inferred from cultured hosts, and it is well known that natural populations harbor a broad diversity of virus and hosts with different levels of susceptibility to infection (Zingone 1999). Alternative methods to calculate the impact of viruses on specific hosts, as transmission electronic microscopy of infected cells, are laborious and time consuming (Zingone *et al.*, 1999).

The advent of molecular techniques opened new venues for studying virus–host interactions in complex communities, such as qPCR detection of virus and host (Sandaa and Larsen, 2006), droplet digital PCR (Lim *et al.*, 2017), single-cell genomics (Roux *et al.*, 2017; Castillo *et al.*, 2019) or correlations between viral genes with their putative hosts found in metagenomes (Mizuno *et al.*, 2013; Nishimura *et al.*, 2017). Although all these approaches have provided very valuable information on marine viral ecology, they do not directly assess the impact of viruses on specific populations. Now, VirusFISH arises as a powerful tool to directly detect targeted viral-host interactions in natural seawater samples, and unlike the other approaches, it allows the visualization of the interaction, and the monitoring of the infection dynamics.

We used VirusFISH to follow the viral-host interactions on *Ostreococcus* populations. Our results showed that *Ostreococcus* displayed in general low abundances, but it occasionally represented up to ~20% of the picoeukaryotes

assemblage (Table S3). This is in agreement with previous results, that showed that *Ostreococcus* is a low abundance picoeukaryote in coastal and offshore waters (<$5{\cdot}10^3$ cells mL^{-1} at surface and DCM) (Zhu *et al.*, 2005; Countway and Caron, 2006; Cardol *et al.*, 2008; Derelle *et al.*, 2015), unlike in lagoons, as for example the Thau Lagoon, where *O. tauri* is generally the main component of the phytoplankton community (Vaquer *et al.*, 1996). However, despite their general low abundances in coastal waters, it has been reported that *Ostreococcus* can produce sporadic blooms, increasing two orders of magnitude its basal concentration and accounting for the 70% of the total picoeukaryotic community (O'Kelly *et al.*, 2003; Countway and Caron, 2006).

We observed a certain seasonal pattern of *Ostreococcus* abundances, corresponding to the absence of cells during the winter season, but viral infection dynamics were variable throughout the year. Yet, we are aware that monthly sampling frequency may not be sufficient to detect episodes of boom and bust in the *Ostreococcus* populations and to quantify the role of viruses in controlling *Ostreococcus* abundance. Using a weekly sampling over three years, it was shown that both *Micromonas pusilla* and its viruses fluctuated widely on small time scales (Zingone et al. 1999). Also, Johannessen *et al.* (2017) reported that Haptophyte and virus communities composition and diversity varied substantially during an annual cycle and uncoordinatedly. These observations suggest that the dynamics of picoeukaryotes in the environment are complex and therefore high frequency samplings should be carried out to address picoeukaryote-virus interactions in nature. Despite this, our work is the first approximation that directly assesses the impact of viruses on a picoeukaryotic population in nature.

With a few exceptions, the highest contribution of *Ostreococcus* to the picoeukaryotic assemblage occurred in the summer (Table S3). This might indicate that this tiny picoeukaryote is better adapted to low nutrient conditions than other members of the picoeukaryotic assemblage in these coastal waters.

The usage of a general *Ostreococcus* CARD-FISH probe does not allow the distinction of specific species. However, metatranscriptomic data unveiled that the dominating species in the surface waters of the continental shelf station was *O. lucimarinus*. A previous study has shown that this species inhabits waters from the surface to the DCM (Rodríguez *et al.*, 2005) and it is the most widely distributed (Tragin and Vaulot, 2019), whereas *O. tauri* and *O. mediterraneus* are mostly restricted to surface waters (Rodríguez *et al.*, 2005). From this information we can infer that likely most *Ostreococcus* cells found in samples from 50m depth belonged to *O. lucimarinus*.

Our VirusFISH probes were designed for the *Ostreococcus* virus OtV5 (Chapter 1, see the material and methods section). However, according to Allers *et al*. (2013), only one probe is enough to visually detect one virus, and the detection efficiency increases with the number of viral probes used. Therefore, despite the similarity of our probes with the different viral genomes in some cases was not 100% , we expect that the mix of the designed probes target all the OV, except OtV6, which as shown by previous studies is evolutionarily distinct (Monier *et al.*, 2017).

The metatranscriptomic data showed that the most transcriptionally active species was *O. lucimarinus*, and that several viruses infecting this species coexisted. *Ostreococcus* viral transcriptional activity was higher in 2011 than in 2012, when we did the VirusFISH analyses, but even in 2011 it was low relative to the transcriptional activity of the hosts (Fig. 5). A recent transcriptomic study on Prasinovirus infection of *Ostreococcus* has shown that the viral attack occurs mostly by night (Derelle *et al.*, 2017), which may explain the low viral transcriptional activity detected. Another plausible explanation is that the high diversity of *Ostreococcus* viruses in nature allows them to propagate in a stable coexistence with their hosts, similar to what it has been suggested for field populations of *Micromonas* (Cottrell and Suttle, 1995). Nevertheless, another metatranscriptomic study in the Baltic Sea has shown high *Ostreococcus* viral transcriptional activity relative to *Ostreococcus* (Zeigler Allen *et al.*, 2017), suggesting that we may have missed a large infection event due to our monthly

sampling frequency. A combination of metatranscriptomics with VirusFISH analyses performed with higher sampling frequency should help to gain a clearer insight into the viral–host dynamics of natural populations of *Ostreococcus*.

In conclusion, VirusFISH arises as a powerful technique to follow the dynamics of hosts and their infecting viruses in nature. It requires the previous knowledge of the viral genome, and preferably the host genome to design the adequate probes (i.e. probes that not target by mistake regions of the host genome that are similar to the virus), as well as the viral DNA to use it as template to synthetize the probes. Thus, it can be easily implemented with any genome sequenced virus–host system available in culture. However, VirusFISH could also be used to find the host of abundant viruses detected through metagenomics, provided there is enough viral DNA template to synthetize the probes. Therefore, VirusFISH opens avenues in viral ecology to tackle the role of viruses in controlling the abundance of key players in marine microbial communities, allowing for the first time to visually quantify the impact on specific host populations.

REFERENCES

Alonso-Sáez, L., Morán, X.A.G., and Clokie, M.R. (2018) Low activity of lytic pelagiphages in coastal marine waters. *ISME J.* **12**: 2100–2102.

Arandia-Gorostidi, N., Huete-Stauffer, T.M., Alonso-Sáez, L., and G. Morán, X.A. (2017) Testing the metabolic theory of ecology with marine bacteria: different temperature sensitivity of major phylogenetic groups during the spring phytoplankton bloom. *Environ. Microbiol.* **19**: 4493–4505.

Baudoux, A.C. (2007) The role of viruses in marine phytoplankton mortality. s.n., p. 148.

Bec, B., Husseini-Ratrema, J., Collos, Y., Souchu, P., and Vaquer, A. (2005) Phytoplankton seasonal dynamics in a Mediterranean coastal lagoon: emphasis on the picoeukaryote community. *J. Plankton Res.* **27**: 881–894.

Brussaard, C.P.D., Noordeloos, A.A.M., Sandaa, R.A., Heldal, M., and Bratbak, G. (2004) Discovery of a dsRNA virus infecting the marine photosynthetic protist *Micromonas pusilla*. *Virology* **319**: 280–291.

Cardol, P., Bailleul, B., Rappaport, F., Derelle, E., Beal, D., Breyton, C., et al. (2008) An original adaptation of photosynthesis in the marine green alga *Ostreococcus*. *Proc. Natl. Acad. Sci.* **105**: 7881–7886.

Castillo, Y.M., Mangot, J., Benites, L.F., Logares, R., Kuronishi, M., Ogata, H., et al. (2019) Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics. *Mol. Ecol.* **28**: 4272–4289.

Chen, F. and Suttle, C.A. (1995) Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Appl. Environ. Microbiol.* **61**: 1274–8.

Cottrell, M. and Suttle, C. (1991) Wide-spread occurrence and clonal variation in viruses which cause lysis of a cosmopolitan, eukaryotic marine phytoplankter *Micromonas pusilla*. *Mar. Ecol. Prog. Ser.* **78**: 1–9.

Cottrell, M.T. and Suttle, C.A. (1995) Dynamics of lytic virus infecting the photosynthetic marine picoflagellate Micromonas pusilla. *Limnol. Oceanogr.* **40**: 730–739.

Countway, P.D. and Caron, D.A. (2006) Abundance and distribution of Ostreococcus sp. in the San Pedro Channel, California, as revealed by quantitative PCR. *Appl. Environ. Microbiol.* **72**: 2496–506.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5**: 1095–1107.

Derelle, E., Ferraz, C., Escande, M.-L., Eychenié, S., Cooke, R., Piganeau, G., et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS One* **3**: e2250.

Derelle, E., Ferraz, C., Rombauts, S., Rouze, P., Worden, A.Z., Robbens, S., et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci.* **103**: 11647–11652.

Derelle, E., Monier, A., Cooke, R., Worden, A.Z., Grimsley, N.H., and Moreau, H.

(2015) Diversity of viruses infecting the green microalga *Ostreococcus lucimarinus*. *J. Virol.* **89**: 5812–5821.

Derelle, E., Yau, S., Moreau, H., and Grimsley, N.H. (2017) Prasinovirus attack of *Ostreococcus* is furtive by day but savage by night. *J. Virol.* **92**: JVI.01703-17.

Field, C.B. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237–240.

Fogg, G.E. and Thake, B. (1987) Algal cultures and phytoplankton ecology. University of Wisconsin Press.

JMP®, Version 9.0.1. SAS Institute Inc., Cary, NC, 1989-2019.

Johannessen, T., Larsen, A., Bratbak, G., Pagarete, A., Edvardsen, B., Egge, E., and Sandaa, R.-A. (2017) Seasonal dynamics of Haptophytes and dsDNA algal viruses suggest complex virus-host relationship. *Viruses* **9**: 84.

Lara, E., Vaqué, D., Sà, E.L., Boras, J.A., Gomes, A., Borrull, E., et al. (2017) Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci. Adv.* **3**: e1602565.

Larsen, J.B., Larsen, A., Bratbak, G., and Sandaa, R.-A. (2008) Phylogenetic analysis of members of the Phycodnaviridae virus family, using amplified fragments of the major capsid protein gene. *Appl. Environ. Microbiol.* **74**: 3048–3057.

Lehahn, Y., Koren, I., Schatz, D., Frada, M., Sheyn, U., Boss, E., et al. (2014) Decoupling physical from biological processes to assess the impact of viruses on a mesoscale algal bloom. *Curr. Biol.* **24**: 2041–2046.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Lim, S.W., Lance, S.T., Stedman, K.M., and Abate, A.R. (2017) PCR-activated cell sorting as a general, cultivation-free method for high-throughput identification and enrichment of virus hosts. *J. Virol. Methods* **242**: 14–21.

Massana, R. (2011) Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**: 91–110.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., and Ghai, R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**: e1003987.

Mojica, K.D.A., Huisman, J., Wilhelm, S.W., and Brussaard, C.P.D. (2016) Latitudinal variation in virus-induced mortality of phytoplankton across the North Atlantic Ocean. *ISME J.* **10**: 500–513.

Monier, A., Chambouvet, A., Milner, D.S., Attah, V., Terrado, R., Lovejoy, C., et al. (2017) Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. *Proc. Natl. Acad. Sci.* **114**: E7489–E7498.

Morán, X.A.G., Calvo-Díaz, A., Arandia-Gorostidi, N., and Huete-Stauffer, T.M. (2018) Temperature sensitivities of microbial plankton net growth rates are seasonally coherent and linked to nutrient availability. *Environ. Microbiol.* **20**: 3798–3810.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., et al. (2017) Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere* **2**: e00359-16.

Not, F., Latasa, M., Marie, D., Cariou, T., Vaulot, D., and Simon, N. (2004) A single

species, Micromonas pusilla (Prasinophyceae), dominates the eukaryotic picoplankton in the Western English Channel. *Appl. Environ. Microbiol.* **70**: 4064–72.

O'Kelly, C.J., Sieracki, M.E., Thier, E.C., and Hobson, I.C. (2003) A transient bloom of *Ostreococcus* (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *J. Phycol.* **39**: 850–854.

R Development Core Team (2016) R: a language and environment for statistical computing.

Redfield, A.C., Ketchum, B.H., and Richards, F.A. (1963) The influence of organisms on the composition of sea-water. In, Hill,M.N. (ed), *The Sea. vol. 2*., pp. 26–77.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.* **29**: 24–26.

Rodríguez, F., Derelle, E., Guillou, L., Le Gall, F., Vaulot, D., and Moreau, H. (2005) Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). *Environ. Microbiol.* **7**: 853–859.

Roux, S., Chan, L.-K., Egan, R., Malmstrom, R.R., McMahon, K.D., and Sullivan, M.B. (2017) Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**: 858.

Sandaa, R.-A. and Larsen, A. (2006) Seasonal variations in virus-host populations in norwegian coastal waters: focusing on the cyanophage ommunity infecting marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **72**: 4610–4618.

Tragin, M. and Vaulot, D. (2019) Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci. Rep.* **9**: 5190.

Vaquer, A., Troussellier, M., Courties, C., and Bibent, B. (1996) Standing stock and dynamics of picophytoplankton in the Thau Lagoon (northwest Mediterranean coast). *Limnol. Oceanogr.* **41**: 1821–1828.

Vardi, A., Haramaty, L., Van Mooy, B.A.S., Fredricks, H.F., Kimmance, S.A., Larsen, A., and Bidle, K.D. (2012) Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proc. Natl. Acad. Sci.* **109**: 19327–19332.

Weynberg, K.D., Allen, M.J., Gilg, I.C., Scanlan, D.J., and Wilson, W.H. (2011) Genome sequence of *Ostreococcus tauri* virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. *J. Virol.* **85**: 4520–4529.

Worden, A.Z., Nolan, J.K., and Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol. Oceanogr.* **49**: 168–179.

Worden, A.Z. and Not, F. (2008) Ecology and Diversity of Picoeukaryotes. In, *Microbial Ecology of the Oceans*. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 159–205.

Yau, S., Hemon, C., Derelle, E., Moreau, H., Piganeau, G., and Grimsley, N. (2016) A viral immunity chromosome in the marine picoeukaryote, Ostreococcus tauri. *PLOS Pathog.* **12**: e1005965.

Yau, S., Krasovec, M., Rombauts, S., Groussin, M., Benites, L.F., Vancaester, E., et al. (2019) Virus-host coexistence in phytoplankton through the genomic lens. *bioRxiv* 513622.

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**: e00125-16.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**: 79–92.

Zingone, A., Sarno, D., and Forlani, G. (1999) Seasonal dynamics in the abundance of Micromonas pusilla (Prasinophyceae) and its viruses in the Gulf of Naples (Mediterranean Sea). *J. Plankton Res.* **21**: 2143–2159.

SUPPLEMENTARY INFORMATION

Supplementary tables

Table S1. Number of reads and average length (in nucleotides) obtained from Illumina Miseq run.

	Raw reads		Quality trimmed reads		Non-rRNA reads
	Number	Length	Number	Length	Number
April 2011	1,833,313	237	1,209,799	152	532,706
May 2011	3,728,913	215	2,599,188	151	1,222,358
July 2011	2,614,931	214	1,795,529	148	564,572
November 2011	4,197,507	218	2,821,587	149	780,955
April 2012	1,385,140	215	990,603	150	372,78
May 2012	1,094,949	209	784,096	150	201,51
July 2012	1,442,785	219	998,929	148	322,105
November 2012	1,535,942	218	1,054,262	150	262,37
TOTAL	17,833,480		12,334,993		4,259,356

Table S2. Biological and physico-chemical parameters of the samples.

Depth	Station	Month	pChl	pChl/Chla total (%)	NO_3	PO_4	NO_3/PO_4
Surface	1	January	0.25	54.5	3.8	0.4	10.6
		February	0.11	49.5	4.6	0.7	6.3
		March	0.29	43.3	–	–	–
		April	0.43	36.4	1.4	0.1	13.6
		May	0.05	17.5	0.03	0.2	0.1
		June	0.07	14.4	0.02	0.2	0.1
		July	0.12	37.5	0.1	0.1	0.6
		August	0.64	39.5	1.8	0.1	20.2
		September	0.36	43.1	0.03	0.1	0.4
		October	0.38	37.3	0.2	0.1	2.2
		November	0.37	34.5	1.7	0.3	5.9
		December	0.78	50.6	4.0	0.2	23.6
Surface	2	January	0.24	50.5	2.3	0.2	10.2
		February	0.18	40.4	4.0	0.1	40.4
		March	0.18	14.1	4.2	0.2	19.1
		April	0.31	31.8	1.3	0.1	12.0
		May	0.03	3.8	0	0.1	0
		June	0.06	26.5	0.1	0.1	1.3
		July	0.06	32.3	0.5	0.1	7.7
		August	0.14	37.8	0.6	0	–
		September	0.15	29.1	0.3	0.1	4.5
		October	0.48	59.1	0.2	0.0	7.0
		November	0.50	47.1	0.6	0.1	9.8
		December	0.59	37.3	3.5	0.2	16.0
Surface	3	January	0.26	53.2	3.6	0.5	7.8
		February	0.11	29.0	3.5	0.1	31.8
		March	0.13	14.7	3.7	0.2	24.8
		April	0.36	38.2	1.2	0.1	13.8
		May	0.06	10.8	0.03	0.2	0.1
		June	0.01	21.2	0.02	0.2	0.1
		July	0.07	38.9	0.03	0.1	0.4
		August	0.07	14.5	1.2	0.1	19.2
		September	0.27	35.0	0.1	0.1	0.7
		October	0.54	48.0	0.03	0.1	0.4
		November	0.34	34.4	0.4	0.2	2.2
		December	0.51	46.8	1.6	0.1	17.7

Table S2. Continuation.

Depth	Station	Month	pChl	pChl/Chla total (%)	NO_3	PO_4	NO_3/PO_4
50 m	2	January	0.12	28.9	4.2	0.2	22.1
		February	0.10	25.2	3.9	0.1	43.2
		March	0.05	23.0	5.7	0.4	13.0
		April	0.23	31.6	1.3	0.2	7.9
		May	0.08	9.2	3.8	0.2	19.0
		June	0.11	19.2	0.4	0.2	2.4
		July	0.28	40.9	1.1	0.2	7.0
		August	0.64	39.0	2.6	0.2	17.5
		September	0.39	30.7	3.7	0.3	12.4
		October	0.36	48.1	0.2	0.03	7.0
		November	0.49	62.7	0.9	0.1	10.2
		December	0.08	16.3	3.2	0.2	13.5
50 m	3	January	0.26	52.6	2.2	0.6	3.9
		February	0.10	35.7	3.5	0.2	17.7
		March	0.07	31.3	5.2	0.4	14.9
		April	0.17	72.8	3.2	0.2	19.8
		May	0.07	4.3	2.1	0.3	7.2
		June	0.05	26.8	0.6	0.2	3.2
		July	0.18	35.5	0.1	0.1	0.9
		August	0.36	28.9	0.7	0.1	8.5
		September	0.52	19.6	0.3	0.1	2.8
		October	0.56	55.3	0.03	0.1	0.4
		November	0.33	48.5	0.7	0.1	6.1
		December	0.44	38.2	1.7	0.3	5.5

Abbreviations: pChl: chlorophyll in the picoplankton size fraction; Chla: total Chlorophyll *a*; "–", No data.

Table S3. Relative contribution (%) of *Ostreococcus* to the total small picoeukaryotic community, at surface and 50 m depth, in 2012.

	Surface			50m depth	
	E1	**E2**	**E3**	**E2**	**E3**
January	0.5	0.6	0.4	ND	ND
February	4	0.6	ND	1.7	0.8
March	ND	ND	1.5	ND	ND
April	ND	ND	ND	1.9	ND
May	1.1	6.7	9.5	2.2	3.1
June	3	1.9	1	8.9	0.2
July	1.3	20.8	12.5	2.2	–
August	8	ND	ND	2.5	1.3
September	0.6	1	2.9	4.8	1.4
October	1.4	ND	1.5	ND	ND
November	0.5	0.9	1.1	1.2	2.1
December	1.5	4.6	–	3.9	–

Abbreviations: E1, station 1; E2, station 2; E3, station 3; ND, Non-detected; "–", No data available.

Table S4. Correlation analyses between all the biological and physico-chemical variables, including *Ostreococcus* and the abundance of infected cells (infection). Only significant correlations (p-value < 0.05) are shown.

Variable 1	Variable 2	R	n	p-value
Picoeuk abundance	Infection	0.43	54	0.0013
Picoeuk abundance	*Ostreo* abundance	0.42	54	0.0016
Chla	Picoeuk abundance	0.29	57	0.027
Chla	pChl	0.75	60	0.0001
NO_3	Picoeuk abundance	-0.31	56	0.02
NO_3	Temperature	-0.69	59	0.0001
pChl	Picoeuk abundance	0.40	57	0.002
PO_4	NO_3	0.58	59	0.0001
PO_4	Temperature	-0.53	59	0.0001
Salinity	Picoeuk abundance	-0.45	57	0.0005
Salinity	Chla	-0.32	60	0.014
Salinity	pChl	-0.31	60	0.016
Salinity	Temperature	0.28	60	0.032
Salinity	Infection	-0.32	57	0.016
Temperature	*Ostreo* abundance	0.29	57	0.027

Abbreviations: *Ostreo*, *Ostreococcus*; pChl, chlorophyll in the pico size fraction; Chla, total chlorophyll *a*; Picoeuk, picoeukaryotes.

Supplementary figures

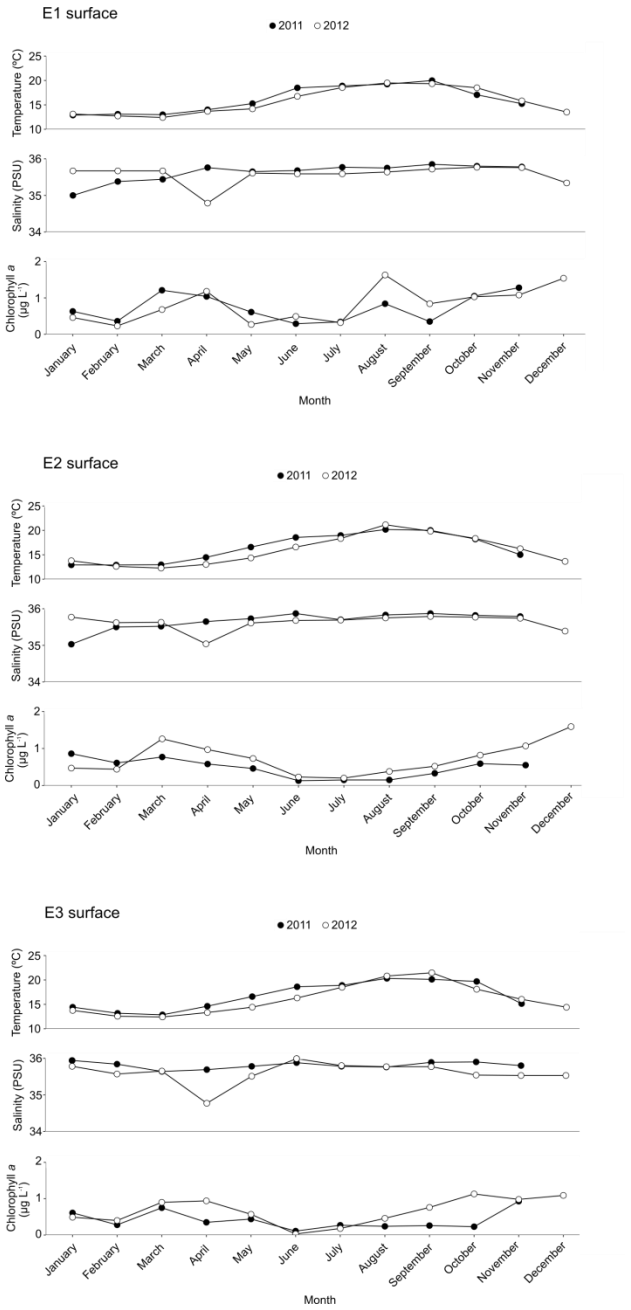


Figure S1. Surface and 50 m depth *in situ* temperature, salinity and chlorophyll *a* concentration in 2011 and 2012 at the coastal (E1), continental shelf (E2) and offshore (E3) sites. The figure continues in the next page.

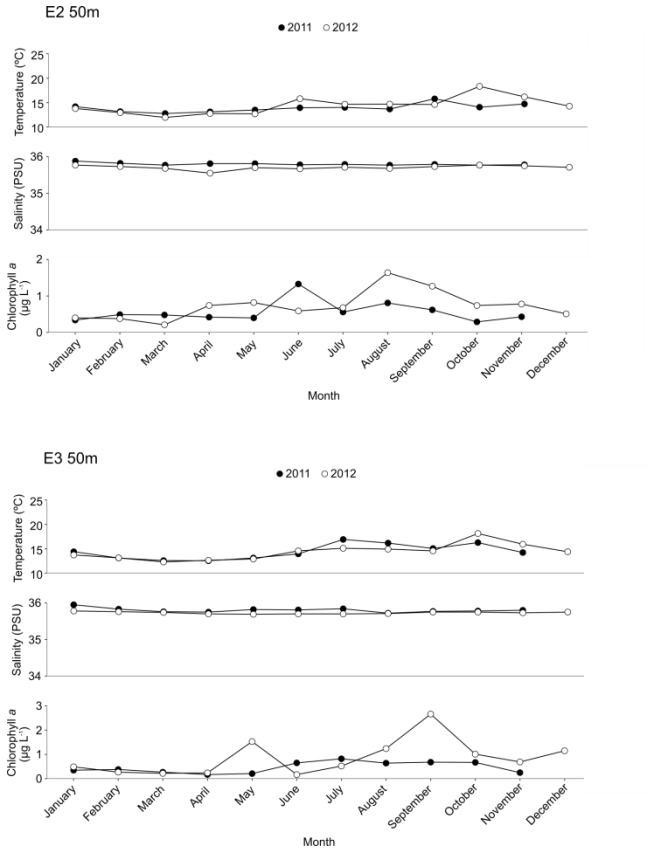
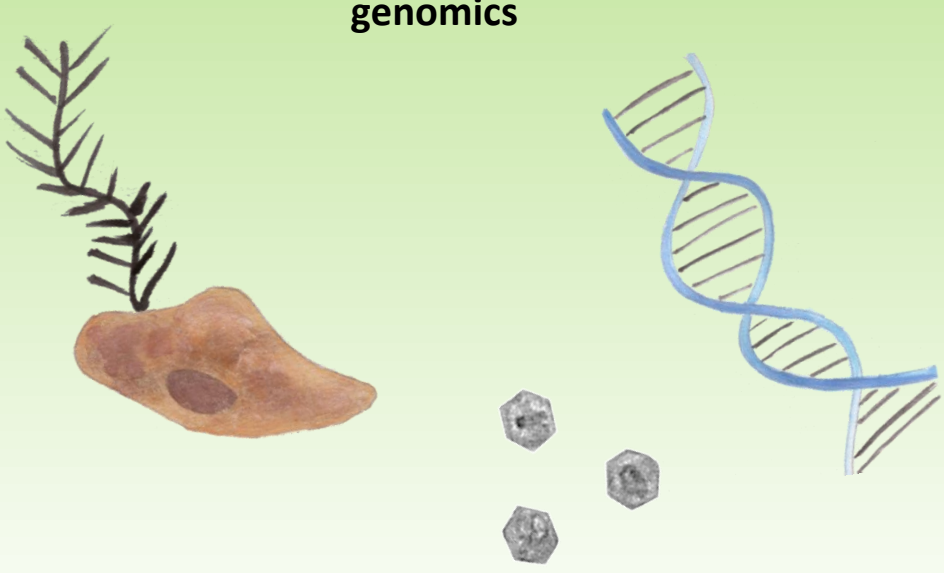


Figure S1. Continuation.

Figure S2. Metatranscriptomic read alignment to the OtV5 genome. Metatranscriptomic reads are shown as red lines. Predicted genes are shown as blue blocks with transcribed genes with functional annotations predicted to be important in viral replication are coloured in magenta. These are from left to right: mRNA capping methyltransferase (mRNA capping). viral type A inclusion protein (Viral inclusion protein). adenine methyltransferase (adenineMT). ribonucleoside reductase small subunit (RNR). transcription initiation factor IIB (TFIIB). major capsid protein (MCP). ATP-dependent protease (protease). dUTPase. DNA polymerase B (DPOB) and topoisomerase I.

CHAPTER 3

Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics



Castillo, Y. M., Mangot, J., Benites, L. F., Logares, R., Kuronishi, M., Ogata, H., Jaillon, O., Massana, R., Sebastián, M. and Vaqué, D. (2019). Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics. *Molecular Ecology*. **28:** 4272–4289. doi:10.1111/mec.15210.

ABSTRACT

Viruses are the most abundant biological entities on Earth and have fundamental ecological roles in controlling microbial communities. Yet, although their diversity is being increasingly explored, little is known about the extent of viral interactions with their protist hosts as most studies are limited to a few cultivated species. Here, we exploit the potential of single-cell genomics to unveil viral associations in 65 individual cells of 11 essentially uncultured stramenopiles lineages sampled during the *Tara* Oceans expedition. We identified viral signals in 57% of the cells, covering nearly every lineage and with narrow host specificity signal. Only 7 out of the 64 detected viruses displayed homologies to known viral sequences. A search for our viral sequences in global ocean metagenomes showed that they were preferentially found at the DCM and within the 0.2-3 µm size fraction. Some of the viral signals were widely distributed, while others geographically constrained. Among the viral signals we detected an endogenous mavirus virophage potentially integrated within the nuclear genome of two distant uncultured stramenopiles. Virophages have been previously reported as a cell's defense mechanism against other viruses, and may therefore play an important ecological role in regulating protist populations. Our results point to single-cell genomics as a powerful tool to investigate viral associations in uncultured protists, suggesting a wide distribution of these relationships, and providing new insights into the global viral diversity.

KEYWORDS

Single-cell genomics; viral associations; protists; uncultured stramenopiles; viruses; virophages

3.1. INTRODUCTION

Viruses are major players in marine biogeochemical cycles (Jover *et al.*, 2014) and constitute the most abundant biological entities in the oceans, ranging from about 10^4 to 10^7 ml^{-1} (Suttle, 2005; Danovaro *et al.*, 2011). They are known to be a major cause of microbial (bacteria, archaea and protists) mortalities (Munn, 2006), leading to approximatively 10^{29} infection events every day (Brussaard et al., 2008) and causing the release of 10^8-10^9 tons of biogenic carbon per day (Brussaard et al., 2008; Suttle, 2005). Furthermore, they are main vectors of gene transfer in the oceans (Middelboe and Brussaard, 2017), impacting microbial community dynamics, diversity and evolution (Breitbart, 2012; Weitz and Wilhelm, 2012; Jover *et al.*, 2014).

Our knowledge of marine viral diversity and biogeography has been constantly expanding during this last decade with the advent of viral metagenomics (e.g., Coutinho et al., 2017; Mizuno, Rodriguez-Valera, Kimes, & Ghai, 2013; Paez-Espino et al., 2016). Multiple studies unveiled a large novel diversity of uncultured viruses, indicating their key roles in nutrient cycling and trophic networks (Brum *et al.*, 2015; Roux *et al.*, 2016). Unfortunately, despite these fruitful advances in viral ecology, our understanding of virus-host interactions is still in its infancy. The question of 'who infects whom' within marine microbial communities has always been central, and the assessment of the true extent of host specificity among marine viruses remains challenging (Brum and Sullivan, 2015). For a long time, studies investigating virus-host interactions were limited to cultured host cells, restricting our knowledge to the 0.1–1% of host cells that are in culture (Rappé and Giovannoni, 2003; Swan *et al.*, 2013), and biasing our knowledge towards virulent lytic viruses (Brüssow and Hendrix, 2002; Swan *et al.*, 2013). Thus, many viruses are still uncharacterized and novel culture-independent approaches are needed to overcome these methodological limitations.

Several methods have been developed to investigate putative interactions between viruses and uncultured hosts as reviewed by Brum & Sullivan (2015) and Breitbart, Bonnain, Malki, & Sawaya (2018). These include analyses by metaviromics (Bolduc *et al.*, 2015; Brum *et al.*, 2015), matching CRISPR spacers (Anderson *et al.*, 2011; Berg Miller *et al.*, 2012), phageFISH (Fluorescence *In Situ* Hybridization; Allers et al., 2013), viral tagging (Deng *et al.*, 2012, 2014), the polony method (Baran *et al.*, 2018), the use of microfluidic digital PCR (Tadmor *et al.*, 2011) and single-cell genomics (SCG) (e.g., Labonté et al., 2015 and Roux et al., 2014). From these, SCG emerged as a powerful complement to cultivation and metagenomics by providing genomic information from individual uncultured cells (Stepanauskas, 2012). Furthermore, it has an incredible potential for cell-specific analyses of organismal interactions, such as parasitism, symbiosis and predation (Stepanauskas, 2012; Krabberød *et al.*, 2017), giving comprehensive insights of *in situ* virus-host associations. Indeed, this effective approach has revealed new associations between viruses and bacterial (Roux *et al.*, 2014; Labonté *et al.*, 2015) or archaeal cells (Chow *et al.*, 2015; Labonté *et al.*, 2015; Munson-McGee *et al.*, 2018). However, the application of SCG to protist cells is relatively recent and there is still a limited number of Single-cell Amplified Genomes (SAGs) from microeukaryotes (e.g., Bhattacharya et al., 2012; Heywood, Sieracki, Bellows, Poulton, & Stepanauskas, 2011; Mangot et al., 2017; Roy et al., 2015; Troell et al., 2016; Vannier et al., 2016), with only one study that has so far explored virus-host interactions (Yoon *et al.*, 2011).

In the present work, we use SCG to uncover putative interactions between viruses and uncultured protists using 65 SAGs produced during the *Tara* Oceans expedition (Karsenti *et al.*, 2011). These cells were affiliated to 11 stramenopile lineages belonging to MArine STramenopiles (MASTs), Chrysophyceae, Dictyochophyceae and Pelagophyceae, that are known to be important components of marine pico- and nanosized eukaryotic assemblages (1-5 µm, Massana, 2011). Initially detected in molecular diversity surveys, MASTs are formed by at least 18 independent groups of essentially uncultured protists

(Massana *et al.*, 2014), some of which display a widespread distribution in sequencing data sets (Lin *et al.*, 2012; Logares *et al.*, 2012; Seeleuthner *et al.*, 2018) and are abundant in microscopy counts (Massana *et al.*, 2006). Within these MASTs, we analyzed SAGs from three clades, MAST-3, MAST-4 and MAST-7 (Massana *et al.*, 2014). We also report the putative linkages between viruses and SAGs from the uncultured chrysophyte lineages G and H, formed by pigmented and colorless cells respectively, which are abundant in molecular diversity surveys (del Campo and Massana, 2011; Seeleuthner *et al.*, 2018). Finally, we screened for viral signatures in SAGs from a cultured pelagophyte (*Pelagomonas calceolata*) and an uncultured dictyochophyte within the order Pedinellales.

Our results revealed a large diversity of viral sequences associated to protist cells, the vast majority of which correspond to previously unidentified viral lineages. Using global ocean metagenomes from the *Tara* Oceans expedition, we looked at the geographical distribution of the identified viral sequences in epipelagic waters by fragment recruitment analysis, finding that some SAG-associated viruses were widely distributed while others were restricted to certain areas. Finally, special attention was paid to a particular virophage sequence retrieved in two distinct stramenopile lineages that is highly similar to the endogenous *Cafeteriavirus*-dependent mavirus, known to be integrated within the nuclear genome of their host (Fischer and Hackl, 2016). Overall, our approach constitutes an initial attempt to determine virus-host associations within protists using culture-independent single-cell genomics.

3.2. MATERIALS AND METHODS

3.2.1. Sample collection and single cell sorting

Samples for single-cell sorting were collected during the circumglobal *Tara* Oceans expedition (2009-2013) (Karsenti *et al.*, 2011) and processed as described in Alberti *et al*., 2017. Flow cytometry cell sorting on cryopreserved samples and genomic DNA amplification by multiple displacement amplification (MDA) were performed at the Single Cell Genomics Center in the Bigelow Laboratory (https://scgc.bigelow.org). SAGs from phototrophic (plastidic) and heterotrophic (aplastidic) cells were screened by PCR using universal eukaryote DNA primers and taxonomically assigned. A total of 65 SAGs affiliated to 11 stramenopiles lineages (Table S1) were selected for sequencing. Sequence data is available at ENA (http://www.ebi.ac.uk/services/tara-oceans-data) under the accession codes listed in Table S2. Main sample-associated environmental data are reported in Table S3, and more details can be found in PANGAEA (Pesant *et al.*, 2015).

3.2.2. SAG sequencing and assembly

After purification of MDA products, 101 bp paired-end libraries were prepared from each single cell as described in Alberti et al., 2017 and cells were independently sequenced on a 1⁄8th Illumina HiSeq lane at the Oregon Health & Science University (US) or at the National Sequencing Center of Genoscope (France). Reads from SAGs were assembled using SPAdes 3.1 (Nurk *et al.*, 2013). In all assemblies, contigs shorter than 500bp were discarded. Quality profiles and basic statistics (genome size, number of contigs, N50, GC content) of each SAG assemblies were generated with Quast (Gurevich *et al.*, 2013). Estimations of genome recovery were done with BUSCO (Benchmarking Universal Single-Copy Orthologs; Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015).

3.2.3. Detection and identification of viral signals in SAGs

Putative viral sequences were retrieved from each assembled SAG using VirSorter v1.0.3 (Roux *et al.*, 2015) with default parameters and both the *RefSeqABVir* and *Virome* databases through the CyVerse Discovery Environment (Devisetty *et al.*, 2016). Contigs identified by VirSorter at all three levels of confidence (from the more to the less confident predictions), categorized as viruses and prophages, were used in subsequent analyses. Sequence similarity between identified full length viral contigs was checked via pairwise BLASTn v2.2.28 (Altschul *et al.*, 1990). Contigs with sequence similarity >95%, coverage >80% and e-value of $<10^{-5}$ were clustered together. Only one representative contig (i.e., the longest one) for each non-redundant SAG-associated viral sequence of each cluster (here after unique contig) was kept for further analysis.

Taxonomy of SAG-associated viral contigs was inferred using the webserver ViPTree (Nishimura *et al.*, 2017). Proteomic trees of each unique contig were generated based on genome-wide sequence similarities computed by tBLASTx. A measure of genomic similarity based on a normalized bit score of tBLASTx (S_G) was calculated against a set of reference viral genomes database, the GenomeNet Virus–Host Database (Mihara *et al.*, 2016). Since MDA does not amplify RNA viruses, only ssDNA and dsDNA viruses and virophage/satellites were considered in that analysis. SAG-associated viruses showing a S_G >0.15 with a reference viral genome were assumed to belong to the same viral genus (Nishimura *et al.*, 2017).

Finally, protein-coding genes of each unique SAG-associated viral sequences were predicted using Prodigal v2.6.3 (Hyatt *et al.*, 2010) and annotated with BLASTp v2.7.1 (e-value 0.001, max. 10 hits) using the NCBI's nr database (updated 09 Feb 2019).

3.2.4. Biogeography of SAG-associated viral contigs assessed by fragment recruitment analysis.

The global distribution of each unique SAG-associated virus was estimated by fragment recruitment analysis against metagenomes from the Ocean Microbial Reference Gene Catalog (OM-RGC; Sunagawa et al., 2015) using an approach similar to Swan et al., 2013. A total of 128 metagenomes from two depths (surface and Deep Chlorophyll Maximum [DCM]) targeting both <0.22 µm (n = 48) and 0.22-3 µm (n= 80) size fractions were analyzed. Metagenomic reads were prior randomly subsampled without replacement to the minimum number of reads within each depth and size fraction using reformat.sh from bbtools suite (https://sourceforge.net/projects/bbmap/). BLAST+ v2.7.1 was then used to recruit reads from the OM-RCG database to each viral sequence (n=64) using default parameter values, except for: -perc_identity 70 -evalue 0.0001. The percentage of unique recruits (~100 bp long and ≥95% identity) from each metagenome matching to each viral sequence was normalized by viral sequence length. SAG-associated viral sequence abundances for each metagenome were calculated from the BLAST output and plotted using custom R scripts.

3.2.5. Identification of virophage contigs and detection/reconstruction of the mavirus integration site

The SAG-associated viral contig SV11, determined by ViPTree in the previous analyses, was highly similar to the virophage genome Maverick-related virus (also referred as mavirus, NC_015230, Fischer & Hackl, 2016), which share an evolutionary origin with a class of self-synthesizing DNA transposons called Maverick/Polinton elements (Fischer and Suttle, 2011). Mavirus was recently found integrated within the nuclear genome of the protist *Cafeteria roenbergensis* in multiple sites, where the endogenous virophage genome (named *Cafeteriavirus*-dependent mavirus and here referred as endogenous mavirus, KU052222) was flanked on either side by terminal inverted repeats (TIRs) (Fischer and Hackl, 2016). However, in comparison with the sequence of

the endogenous mavirus, SV11 virophage sequence was partially incomplete. To determine if the incomplete SAG-associated virophages were potentially integrated within their respective host genome, we proceeded as follows. We first identified putative virophage contigs that could have not been detected with VirSorter because this automated tool was only applied on contigs >500bp, and requires a minimum of two predicted genes per contig to identify it as viral. Consequently, for each SAG containing an associated putative virophage sequence (within chrysophyte-G1 and MAST-3A), all contigs (including fragments <500 bp) were searched by a BLASTn analysis against a manually curated sequence of the endogenous mavirus including the TIRs sequences. For each SAG, contigs with a minimum similarity of 95% and maximal e-value of 10^{-4} with the curated endogenous mavirus genome were assumed to belong to the virophage genome and were aligned to the primarily detected virophage contig using ClustalW (Larkin *et al.*, 2007) as implemented in the Geneious package v10.2.2 (Kearse *et al.*, 2012). Then, to increase the completeness of the SAG-associated virophage genome, a fragment recruitment analysis was performed using BLASTn against all identified virophage contigs in each SAG and reads with at least 99% identity and a maximal e-value of 10^{-4} were kept and assembled to the virophage genome using the Geneious *de novo* assembler with a minimum overlap of 50bp and a minimum identity of 95% (Kearse *et al.*, 2012). Gene prediction of the obtained SAG-associated virophage assemblies was done using Prodigal v2.6.3 and annotated with BLASTp v2.2.28 (e-value 0.001, max. 10 hits) against NCBI's nr database (updated 06 Jun 2017).

3.2.6. Phylogenetic and comparative genomic analysis of the virophages

Phylogenetic analyses of the new SAG-associated virophages were performed with a set of reference virophage sequences from the literature. These include the virophage sequences isolated from cultures such as Sputnik (La Scola *et al.*, 2008), Sputnik 2 (Desnues *et al.*, 2012), Sputnik 3 (Gaia *et al.*, 2013), Zamilon

(Gaia *et al.*, 2014) and mavirus (Fischer and Suttle, 2011), combined with sequences assembled from environmental surveys such as Yellowstone Lake (YLSV1-4 (Zhou *et al.*, 2013) and YLSV5-7 (Zhou *et al.*, 2015)), Qinghai Lake (QLV, (Oh *et al.*, 2016)), Dishui Lake (Dishui, (Gong *et al.*, 2016)), Organic Lake (OLV, (Yau *et al.*, 2011)), Ace Lake (ALM, (Zhou *et al.*, 2013)), Trout Bog epilimnion and hypolimnion (TBE and TBH, Roux et al., 2017) and Mendota (Roux *et al.*, 2017).

As proposed by Roux *et al*., 2017, phylogenetic trees were built based on a concatenated alignment using four core genes (major capsid protein [MCP], minor capsid protein [mCP], DNA packaging enzyme [ATPase], and cysteine protease [CysProt]) from all virophage genomes, except for the virophage TBE_1002136, which lacked the ATPase. For this last, only 3 genes were included in the multi-marker alignment. For each virophage core gene, individual alignments were generated with MAFFT v7.305b (L-INS-I algorithm, (Katoh and Standley, 2013)), automatically curated to remove all non-informative positions using trimAl v1.2 (Capella-Gutierrez *et al.*, 2009) and evaluated for optimal amino acid substitution models using ProtTest v3.4.2 (Darriba *et al.*, 2011). The concatenation of the four core genes alignments was performed using a supermatrix approach with a custom python script (https://github.com/wrf/supermatrix/blob/master/add_taxa_to_align.py).

Maximum-likelihood trees of each four individual core genes alignments and the concatenated alignment were constructed with RAxML v. 8.2.9 (Stamatakis, 2014) with 100 trees for both topology and rapid bootstrap analyses, and using the evolutionary models LG+I+G+F (ATPase, CysProt and mCP) and RtREV+I+G+F (MCP).Trees were generated using the ape (Paradis *et al.*, 2004) and ggtree packages (Yu *et al.*, 2017) in R 3.5.1. (R Development Core Team, 2016), and rooted using QLV. To verify the topology of the trees, bayesian phylogenies on each alignment were also generated with MrBayes v3.2.6 (2,000,000 generations; Ronquist *et al*., 2012).

Finally, whole-genome synteny comparisons between chrysophyte-G1 and MAST-3A SAG-associated virophages and their closest published relatives (endogenous *Cafeteriavirus*-dependent mavirus and Ace Lake mavirus) were performed with EasyFig v.2.2.2 (Sullivan *et al.*, 2011) using tBLASTx and filtering of small hits and annotations option. Since all chrysophyte-G1 SAG-associated virophage are highly similar (mean identity of 99%), only one representative sequence (i.e., longest assembly) per stramenopile lineage are displayed (AB233-L11 for chrysophyte-G1 and AA240-G22 for MAST-3A).

3.3. RESULTS

3.3.1. Detection of viral contigs in protist SAGs

We used a total of 65 SAGs from photosynthetic and heterotrophic stramenopiles selected from four *Tara* Oceans stations located in the Mediterranean Sea and Indian Ocean (Table S2): 6 from two lineages of MAST-3 (clades A and F), 27 from three MAST-4 lineages (clades A, C and E), 6 from a lineage of MAST-7 (clade A), 15 from three lineages of Chrysophyceae (clades G1, H1 and H2), 4 from an uncultured clade of Dictyochophyceae and 7 affiliated to *Pelagomonas calceolata* (Table 1, Table S1). Using a relatively similar sequencing depth (mean of 4.99 ± 0.81 Gbp), assembly sizes were variable among the SAGs, averaging from 3.6 (± 2.8) Mbp in Dictyochophyceae to 11.0 (± 8.0) Mbp in MAST-3F (considering contigs >500 bp; Table 1). The variation in assembly completeness was also important, ranging on average from about 1% (in Pelagophyceae and Dictyochophyceae) to 10% (in MAST-3, MAST-4 and chrysophyte-G1; Table 1). Finally, the number of contigs assembled and their respective N50 also varied among SAGs and stramenopile lineages (Table 1).

We first investigated the presence of viral contigs in the 65 stramenopile SAGs assemblies, identifying a total of 79 putative viral sequences in 37 SAGs (~57%) distributed among most analyzed lineages, with the exception of *Pelagomonas calceolata* (Fig. 1a, Table S1). Only two lineages (MAST-4C and MAST-4E) showed

less than half of their cells harboring viral contigs (Fig. 1a, Table S1). Interestingly, a significant fraction of the SAG-associated viruses (~50%) was found in cells affiliated to chrysophytes (8, 24 and 12 viral contigs in chrysophyte-G1, -H1 and -H2, respectively; Fig. 1a) with only one cell without any viral sequence detected out of the 15 analyzed cells (Fig 1a). Furthermore, the chrysophyte cells were very rich in viral sequences, with up to 9 viral contigs retrieved in a single SAG of chrysophyte-H1 (AA538_K19; Table S1). However, this was an exception, since in general we detected from 1 to 3 viral contigs per cell (Table S1).

We next explored the uniqueness of the detected viral contigs based on a pairwise comparison of their full-length sequences. Of the 79 viral sequences initially identified, we determined 64 non-redundant (i.e., unique) sequences (Table 2), ranging from 1 to 48.5 kbp in length (median = 5.7 kbp; Table 2). From the 64, 61 were associated to a single stramenopile lineage (~95%), and only 3 viral sequences (~5%) were either shared between two (SV11 and SV28) or four lineages (SV2) (Fig. 1b): SV2 was found in MAST-4 (clades A and E), MAST-7 and chrysophyte-H1, whereas SV11 was detected in chrysophyte-G1 and MAST-3A, and SV28 in MAST-4 clades A and E (Table 2). With respect to the 61 viral contigs present in only one specific lineage, about 98% of them were reported in only one specific cell (Fig. 1b), with the exceptional case of SV51, present in triplicate in the same chrysophyte-H1 cell (AA538_K19, Table S1). Only one viral contig (SV13) was found in two different chrysophyte-H2 cells (Fig. 1b, Table 2).

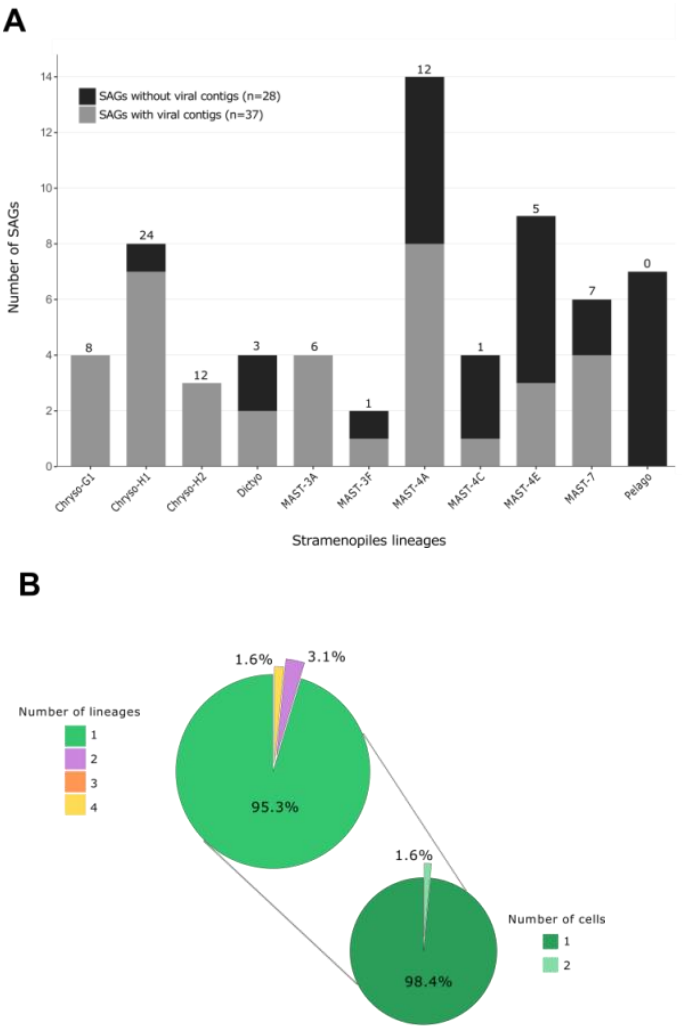


Figure 1. Occurrence and specificity of viral contigs in 65 marine stramenopiles SAGs. **A** Barplots show the number of SAGs with or without any viral contig detected in their assembly. For each lineage, the total number of SAG-associated viral contigs retrieved in SAGs are indicated on top of each bar. **B** Pie charts display the percentage of viral contigs present in only one or shared among 2 or 4 lineages (upper left corner), and the percentage of SAGs that shared a viral contig for those that were lineage-specific (lower right corner). Chryso-G1, Chryso-H1, Chryso-H2, Dictyo and Pelago correspond to the chrysophyte clades G1, H1 and H2, Dictyochophyceae and *Pelagomonas calceolata*, respectively.

Table 1. General characteristics (mean (± standard error)) of the 65 draft stramenopiles SAGs obtained by single-cell genomics.

Group	Name	Number of cells	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)
Chrysophyceae	Chrysophyte-G1	4	5.5 (± 0.5)	9.3 (± 5.2)	3,597 (± 2,009)	11.6 (± 8.2)	40.2 (± 0.2)	5.1 (± 0.4)
	Chrysophyte-H1†	8	4 (± 0.7)	4.0 (± 2.1)	1,425 (± 518)	6.1 (± 3.4)	45.1 (± 0.8)	8.6 (± 3.4)
	Chrysophyte-H2†	3	4.2 (± 1.0)	4.3 (± 2.4)	1,928 (± 1,073)	3.3 (± 1.9)	47.7 (± 1.7)	4.1 (± 0.2)
Dictyochophyceae	unc. dictyochophyte	4	4.6 (± 0.1)	3.6 (± 2.8)	15,567 (± 948)	1.0 (± 1.0)	46.8 (± 2.5)	4.2 (± 1.4)
MAST-3	MAST-3A	4	5.1 (± 0.5)	7.5 (± 1.9)	2,272 (± 409)	11.4 (± 3.4)	42.5 (± 0.3)	8.6 (± 1.1)
	MAST-3F	2	5.4 (± 1.1)	11.0 (± 8.0)	3,576 (± 2,414)	11.7 (± 9.8)	34.1 (± 0.3)	7.7 (± 0.2)
MAST-4	MAST-4A	14	5 (± 1.8)	10.2 (± 4.9)	3,195 (± 1,393)	12.6 (± 7.5)	33.0 (± 1.0)	9.3 (± 2.7)
	MAST-4C	4	5.4 (± 0.6)	8.3 (± 2.3)	2,389 (± 579)	11.8 (± 3.8)	40.3 (± 0.2)	14.0 (± 1.3)
	MAST-4E	9	4.7 (± 0.8)	6.7 (± 2.6)	1,928 (± 607)	8.6 (± 3.9)	44.0 (± 0.7)	9.3 (± 1.9)
MAST-7	MAST-7A	6	5.5 (± 1.3)	5.6 (± 3.2)	2,002 (± 1,233)	3.8 (± 2.1)	44.7 (± 4.8)	7.0 (± 3.2)
Pelagophyceae	*Pelagomonas calceolata*	7	5.6 (± 0.5)	8.1 (± 0.8)	271 (± 186)	0.5 (± 1.0)	47.5 (± 7.6)	13.0 ± (10.9)

Abbreviations: SAG, Single Amplified Genome; MAST, Marine Stramenopiles; unc., uncultured; BUSCO, Benchmarking Universal Single-Copy Orthologs; N50, length of the shortest contig from the minimal set of contig representing 50% of the assembly size.

Table 2. Summary and taxonomic assigment of the 64 SAG-associated viral contigs.

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara *et al*., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV1	MAST-4A (1)	48.5	44	*Cellulophaga* phage (KC821612)	*Podoviridae*	Bacteroidetes	0.06	40.8
SV2	MAST-4A (1), MAST-4E (1), MAST-7 (4), Chryso-H1 (1)	22.8	48	*Prochlorococcus* phage (NC_006883)	*Myoviridae*	Cyanobacteria	0.07	43.2
SV3	MAST-4A (1)	22.1	25	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	< 0.01	31.9
SV4	Chryso-H2 (1)	21.2	25	*Synechococcus* phage (NC_026928)	*Myoviridae*	Cyanobacteria	< 0.01	42.0
SV5	Chryso-H1 (1)	20.5	20	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.04	42.4
SV6	MAST-3A (1)	18.9	19	*Paramecium bursaria Chlorella* virus (NC_009898)	*Phycodnaviridae*	Ciliophora	0.04	39.2
SV7	Chryso-H1 (1)	17.9	32	*Pseudomonas* phage (NC_028980)	*Siphoviridae*	Gammaproteobacteria	0.06	58.7
SV8	MAST-4E (1)	16.7	19	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.01	42.3
SV9	Chryso-H1 (1)	16.7	22	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.07	41.5
SV10	Chryso-H1 (1)	16.2	19	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	40.0
SV11	**Chryso-G1 (4), MAST-3A (1)**	15.5	**18**	**Maverick-related virus (mavirus, NC_015230)**	***Lavidaviridae***	**Bicosoecophyceae**	**0.96**	**98.3**
SV12	MAST-4C (1)	15.0	14	*Rhodothermus* phage (NC_004735)	*Myoviridae*	Bacteroidetes	0.02	39.6

Table 2. Continuation.

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara *et al*., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV13	Chryso-H2 (2)	14.3	11	*Chrysochromulina ericina* virus (NC_028094)	*Phycodnaviridae*	Haptophyta	0.04	66.6
SV14	MAST-4A (1)	13.1	14	-	-	-	-	-
SV15	MAST-4A (1)	12.7	11	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.02	37.6
SV16	MAST-3A (1)	12.6	18	Yellowstone lake phycodnavirus (NC_028110)	*Phycodnaviridae*	*N/D*	0.09	52.6
SV17	Chryso-G1 (1)	10.2	4	*Mycobacterium* phage (NC_028662)	*Podoviridae*	Actinobacteria	0.03	38.5
SV18	Chryso-H1 (1)	10.0	10	*Bacillus* phage (NC_006945)	*Tectiviridae*	Firmicutes	< 0.01	36.1
SV19	MAST-4E (1)	9.9	5	*Vibrio* phage (NC_021529)	*Myoviridae*	Gammaproteobacteria	0.06	49.0
SV20	MAST-7 (1)	9.8	8	*Cronobacter* phage (NC_019398)	*Myoviridae*	Gammaproteobacteria	0.02	51.2
SV21	Chryso-H1 (1)	8.6	8	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.02	35.3
SV22	Chryso-G1 (1)	7.5	11	*Synechococcus* phage (NC_015286)	*Myoviridae*	Cyanobacteria	0.05	49.7
SV23	MAST-7 (1)	7.4	7	*Acanthocystis turfacea Chlorella* virus (NC_008724)	*Phycodnaviridae*	Chlorophyta	0.08	54.9
SV24	Chryso-H1 (1)	7.2	5	-	-	-	-	-
SV25	Dictyo (1)	6.8	8	*Pseudomonas* phage (NC_026600)	*Myoviridae*	Gammaproteobacteria	0.01	43.1
SV26	Chryso-H1 (1)	6.7	9	-	-	-	-	-
SV27	**Chryso-H2 (1)**	6.4	**8**	**Aureococcus anophagefferens virus (NC_024697)**	***Phycodnaviridae***	**Pelagophyceae**	**1.0**	**52.4**

Table 2. Continuation.

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara *et al*., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV28	MAST-4A (2), MAST-4E (2)	5.9	6	*Cellulophaga* phage (KC821612)	*Podoviridae*	Bacteroidetes	0.07	44.2
SV29	Chryso-H1 (1)	5.8	7	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.14	41.0
SV30	Dictyo (1)	5.8	8	*Ostreococcus tauri* virus (NC_010191)	*Phycodnaviridae*	Chlorophyta	0.02	42.9
SV31	Chryso-H1 (1)	5.8	10	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	40.9
SV32	MAST-3A (1)	5.7	6	*Phaeocystis globosa* virus (NC_021312)	*Phycodnaviridae*	Haptophyta	0.07	45.4
SV33	Chryso-G1 (1)	5.6	3	*Anomala cuprea* entomopoxvirus (NC_023426)	*Poxviridae*	Arthropoda	0.1	41.5
SV34	MAST-4A (1)	5.6	2	*Aureococcus anophagefferens* virus (NC_024697)	*Phycodnaviridae*	Pelagophyceae	< 0.01	39.1
SV35	**MAST-3F (1)**	5.4	**7**	**Yellowstone lake phycodnavirus (NC_028110)**	***Phycodnaviridae***	***N/D***	**0.3**	**52.9**
SV36	Chryso-H1 (1)	5.2	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.07	42.5
SV37	Chryso-H1 (1)	5.1	10	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.08	39.8
SV38	Chryso-H2 (1)	4.7	7	*Enterobacteria* phage (NC_005066)	*Myoviridae*	Gammaproteobacteria	0.02	39.7
SV39	Chryso-H1 (1)	4.6	3	*Phaeocystis globosa* virus virophage (NC_021333)	Unclassified virophage	Haptophyta	0.04	35.0
SV40	Dictyo (1)	4.6	4	*Enterobacteria* phage (NC_019526)	*Myoviridae*	Gammaproteobacteria	0.13	44.6
SV41	Chryso-H2 (1)	4.4	7	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	0.04	41.3

Table 2. Continuation.

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara *et al*., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV42	MAST-3A (1)	4.3	6	*Campylobacter* phage (NC_027997)	*Myoviridae*	Epsilonproteobacteria	0.02	30.3
SV43	Chryso-H1 (1)	4.1	6	YSLV7 (NC_028257)	Unclassified virophage	*N/D*	0.04	41.1
SV44	MAST-3A (1)	4.1	5	*Aureococcus anophagefferens* virus (NC_024697)	*Phycodnaviridae*	Pelagophyceae	0.03	32.8
SV45	MAST-4A (1)	3.7	3	*Erwinia* phage (HQ728263)	*Myoviridae*	Gammaproteobacteria	0.02	40.4
SV46	**Chryso-H1 (1)**	3.7	**3**	**YSLV6 (NC_028270)**	**Unclassified virophage**	***N/D***	**0.17**	**40.3**
SV47	Chryso-H1 (1)	3.1	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.09	42.3
SV48	**Chryso-G1 (1)**	3.0	**2**	***Escherichia* phage (NC_025447)**	***Myoviridae***	**Gammaproteobacteria**	**0.2**	**41.4**
SV49	Chryso-H2 (1)	3.0	4	-	-	-	-	-
SV50	**MAST-7 (1)**	2.9	**3**	***Ectocarpus siliculosus* virus (NC_002687)**	***Phycodnaviridae***	**Phaeophyceae**	**0.2**	**46.1**
SV51	Chryso-H1 (1)	2.9	5	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.1	46.5
SV52	Chryso-H1 (1)	2.9	4	-	-	-	-	-
SV53	Chryso-H2 (1)	2.8	4	-	-	-	-	-
SV54	Chryso-H2 (1)	2.8	4	-	-	-	-	-
SV55	Chryso-H2 (1)	2.7	4	-	-	-	-	-
SV56	Chryso-H1 (1)	2.5	4	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.04	42.3

Table 2. Continuation.

SAG-associated viral contig[†]				Taxonomic assigment on the GenomeNet Virus–Host Database (Mihara *et al*., 2016)				
Viral contig	*SAG lineage (number of SAGs)*	*Sequence length (kbp)*	*Number of genes*	*Best viral group hit (GenBank accession number)*	*Viral family*	*Known host group*	*SG*[‡]	*Similarity (%)*
SV57	MAST-4A (1)	2.5	4	*Synechococcus* phage (NC_015289)	*Myoviridae*	Cyanobacteria	0.03	31.1
SV58	MAST-7 (1)	2.3	3	*Enterobacteria* phage (NC_012740)	*Myoviridae*	Gammaproteobacteria	0.05	44.9
SV59	Chryso-H1 (1)	2.3	4	YSLV7 (NC_028257)	Unclassified virophage	*N/D*	0.09	43.3
SV60	Chryso-H1 (1)	2.1	4	YSLV6 (NC_028270)	Unclassified virophage	*N/D*	0.06	43.8
SV61	MAST-4A (1)	2.0	4	-	-	-	-	-
SV62	Chryso-H2 (1)	2.0	4	*Microcystis* phage (NC_029002)	*Myoviridae*	Cyanobacteria	0.04	40.4
SV63	Chryso-H1 (1)	1.8	4	YSLV5 (NC_028269)	Unclassified virophage	*N/D*	0.1	47.0
SV64	**MAST-4A (1)**	1.0	**3**	***Planktothrix* phage (NC_016564)**	***Podoviridae***	**Cyanobacteria**	**0.2**	**49.1**

Abbreviations: SAG, Single Amplified Genome; MAST, Marine Stramenopiles; Chryso, Chrysophyte; Dictyo, Dictyochophyceae; YSLV, Yellowstone Lake virophage. [†]Statistics are computed on the longuest sequence when SAG-associated viral sequences were retrieved in several cell.

[‡] *SG* tBLASTx score. In bold, SAG-associated viral sequence that can be affiliated to the same genus level than their reference best hit (SG > 0.15).

N/D Non Determined. Sequence were isolated from environmental surveys.

Viral signal sequence without taxonomic assignment are shown by the symbol (-).

3.3.2. Diversity and distribution of the SAG-associated viral sequences across the sunlit ocean

The 64 unique viral sequences were compared with a set of reference viral genomes (Mihara *et al.*, 2016). On the basis of high genomic sequence similarity (S_G > 0.15), 7 of the 64 viruses identified in the SAGs could be putatively assigned to four different viral families. These viruses were two virophages (SV11, SV46), three viruses of *Phycodnaviridae* (SV27, SV35, SV50), one virus of *Myoviridae* (SV48) and one virus of *Podoviridae* (SV64; Table 2). Other viruses showed lower sequence similarities (n = 48; S_G < 0.15) or lacked detectable similarity by tBLASTn (n=9) to reference viral genomes, thus being uncertain for their classification at the genus level (Table 2). None of the assigned viral genomes were complete (or circular) but one particular virus, SV11, which seemed nearly complete based on the similarity to a reference genome. This virus of 15.5 kbp in length was highly similar (98.3%, S_G = 0.96) to the Maverick-related virus genome (mavirus, GenBank accession number: NC_015230) and likely belong to the virophage genus of *Mavirus* (*Lavidaviridae*; Table 2). The remaining identified viral signals includes a set of short genome fragments (from 1 to 6.4 kbp) with intermediate genomic similarities (40-53%) to either an unclassified virophage (SV46 with YLV6), eukaryotic viruses (SV27 and SV50 with *Phycodnaviridae*), or phages (SV48 and SV64) (Table 2). We further predicted protein-coding genes in the 64 unique viral sequences. Of the total of 619 predicted genes (median = 6 predicted genes per SV; Table 2), about ~60% (n= 363) had a close relative in the NCBI's nr database and 103 genes were related to eukaryotic viral functions (Table S4).

In order to address the occurrence of these putative viruses in marine epipelagic waters, we performed a fragment recruitment analysis of the viral signals in the *Tara* Oceans OM-RGC database (Sunagawa *et al.*, 2015). Our findings showed that the viral contigs were found preferentially at the DCM, and the 0.2-3 µm size fraction rather than in the <0.2 µm size fraction (Fig. 2 and Fig. S2). Regarding their geographic distribution, the SAG-associated viruses displayed

some differences. On the one hand, some of them showed a cosmopolitan distribution with different degrees of occurrence. For example, some viral contigs (SV1 and SV2) show a high presence in all oceanic basins and in both size fractions, while others (e.g., SV34) were highly present in the 0.2-3 µm size fraction but absent from the <0.2 µm size fraction (Fig. 2). On the other hand, other SAG-associated viruses appeared to be constrained to a lower number of oceanic basins, with some of them showing some biogeography preferences (e.g., SV16 and SV32), whereas others were restricted to few locations with a low presence (e.g., SV53, SV54 and SV55) (Fig. 2 and Fig. S2).

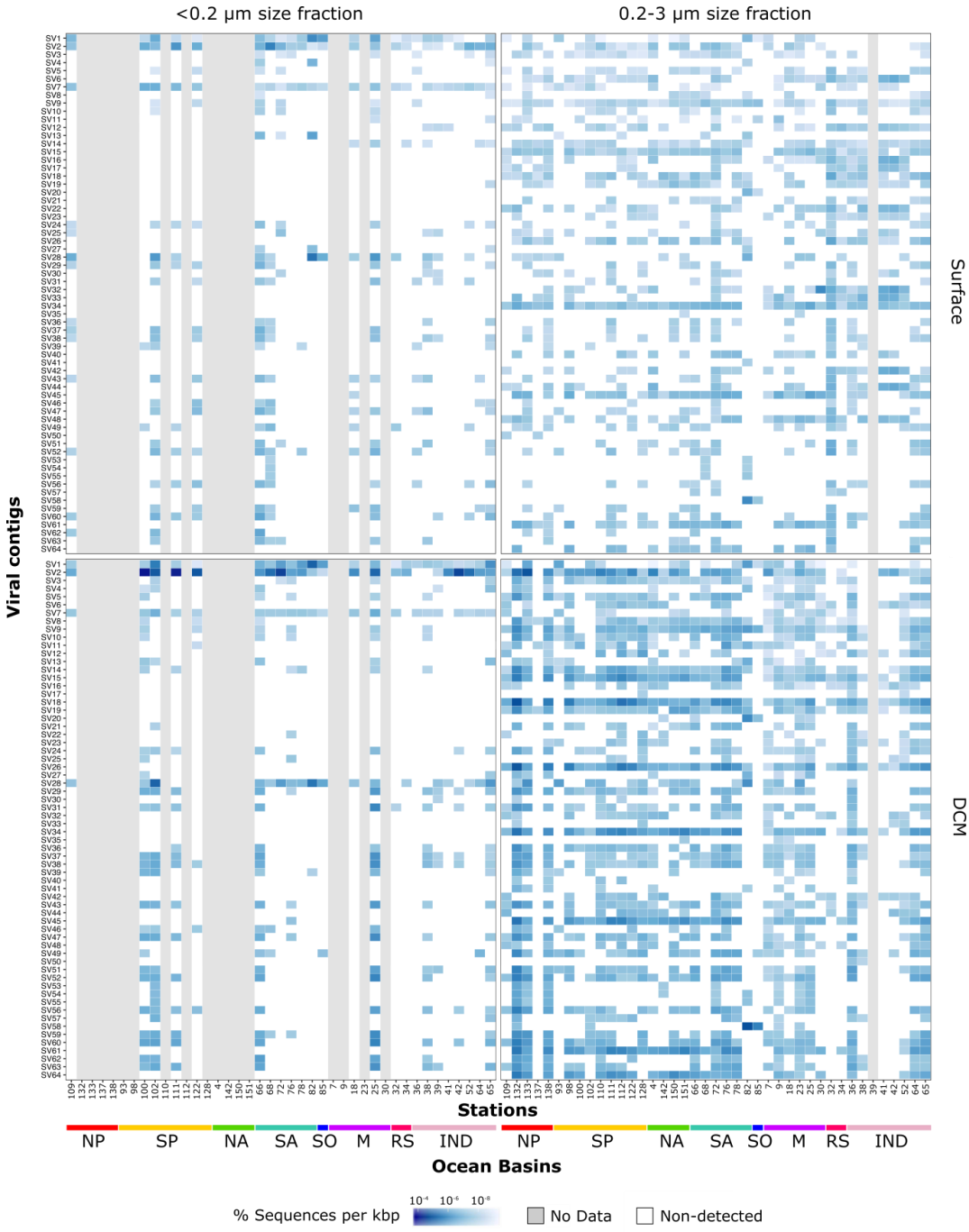


Figure 2. Biogeographical distribution of SAG-associated viruses, as determined by metagenomic fragment recruitment. Viral contigs are shown in the *y*-axis and epipelagic metagenome stations along the *x*-axis. The scale bar indicates the percentage of read

sequences recruited normalized of aligned metagenome sequences with alignments ≥100 bp long and ≥95% identity, normalized by the length of each SAG-associated virus sequence. Results for metagenomes from the <0.2 µm (left panels) and 0.2-3 µm size fractions (right panels) were displayed for both surface (upper panels) and DCM stations (lower panels). Stations where metagenomes were not available are shown in grey (No Data). Color bars represent the different oceanic basins, abbreviations: North Pacific Ocean (NP), South Pacific Ocean (SP), North Atlantic Ocean (NA), South Atlantic Ocean (SA), Southern Ocean (SO), Mediterranean Sea (M), Read Sea (RS) and Indian Ocean (IND).

3.3.3. Genome reconstruction and phylogenetic analysis of SAG-associated virophages

We next focused on five SAG-associated viral contigs (the non-redundant SAG-associated viral contig SV11), retrieved from one MAST-3A (AB240-G22) and from four different chrysophyte-G1 cells (AB233-D06, AB233-L11, AB233-O05 and AB233-P23; Table 2 and Table S1), which were highly similar to mavirus (i.e., Maverick-related virus; Table 2), an endogenous virophage ("provirophage") integrated in the genome of *Cafeteria roenbergensis*. To the best of our knowledge, mavirus constitutes the only case of integration of a *Mavirus* virophage in a protist genome revealed to date by a culture-based approach. However, the virophage genomes identified in each SAG were incomplete compared to mavirus, noting the remarkable lack of two genes coding for an integrase and an helicase, as well as the TIRs, which indicate genome linearity and, therefore, a potential integration into the host genome (Fischer and Hackl, 2016; Roux *et al.*, 2017). After the identification of the putative virophage contigs and a read recruitment analysis within each SAG, to increase the completeness of the SAG-associated virophage genomes (see methods section), we were able to reconstruct the entire SAG-associated virophage genomes of the five stramenopiles cells. This includes the presence of both DNA replication genes and TIRs on either side of all SAG-associated virophage genomes, confirming that SAG-associated mavirus genomes were linear and potentially inserted in the stramenopile host genomes. For the reassembly, from 5 (AB233-

O05) to 14 contigs (AB240-G22), ranging from 0.2-0.3 to 7.2-15.5 kbp in length, were necessary to reconstruct the 5 SAG-associated virophage genomes.

To better assess the phylogenetic position of these newly identified SAG-associated mavirus genomes among the virophages, we established a concatenated marker tree using four virophage core genes (mCP, MCP, ATPase and CysProt; Fig 3), including all the available virophage genomes retrieved from culture, metagenomes and the five new SAG-associated virophages. We found that the newly identified virophage sequences form a clade among the genus *Mavirus* together with the mavirus virophage but distinct from the Ace Lake mavirus (ALM (Zhou *et al.*, 2013), a partial *Mavirus* genome retrieved from an environmental sequencing survey) (Fig. 3). Similar phylogenetic placements were found when each core gene was analyzed separately (Fig. S1).

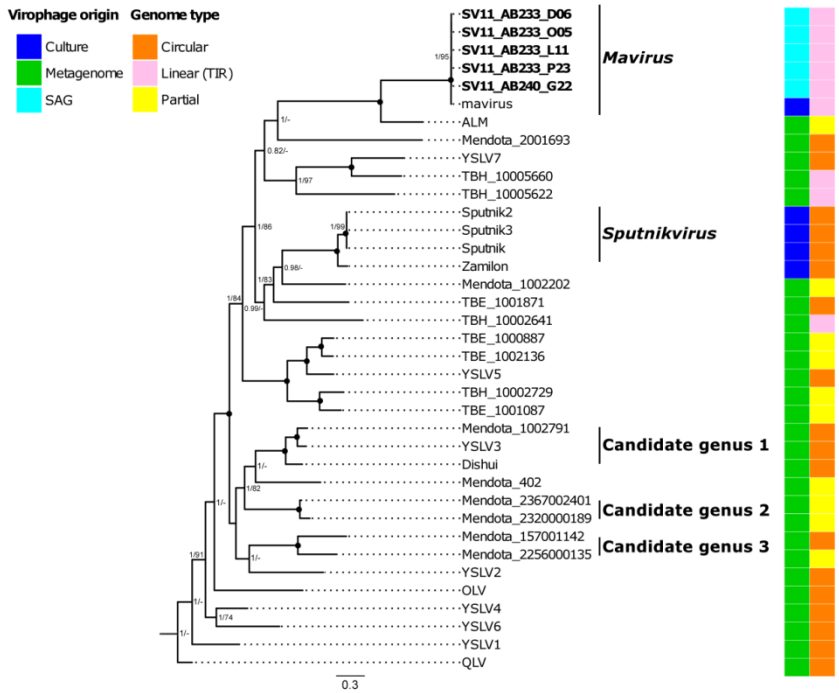


Figure 3. Phylogenetic placement of the new putative SAG-associated mavirus among virophages. The tree topology was inferred from a maximum-likelihood analysis of a concatenated alignment of four core genes (minor [mCP] and major [MCP] capsids proteins, DNA packaging enzyme [ATPase] and Cysteine Protease [CysProt]). Bayesian

posterior probabilities (BPP) and bootstrap percentages (BS) are provided at each node (BPP/BS) when support values were higher than 0.7 and 70%, respectively. Black dots indicate maximal support for both posterior probabilities (1.0) and maximum-likelihood bootstraps (100%) at the respective nodes. The five new SAG-associated virophages are highlighted in bold. The origin (culture, metagenome sequencing or SAG) and genome type (linear with TIRs, circular or partial) of each virophage genome are pointed out in the tree. Abbreviated names for virophages are detailed in the Materials and Methods section.

Finally, we compared the general genome organization of the identified SAG-associated mavirus in MAST-3A (SV11_AB240_G22) and chrysophyte-G1 (SV11_AB233_L11) and their closest published relatives, the endogenous mavirus and Ace Lake mavirus. As expected from the previous analysis, the two SAG-associated mavirus displayed remarkable sequence similarity with the endogenous mavirus integrated within the nuclear genome of *Cafeteria roenbergensis* and exhibited clear differences with the Ace Lake mavirus (Fig. 4). The main differences between the two SAG-associated mavirus and the endogenous mavirus are the presence of an extra gene coding for an unknown function (gene 11, 71 amino acids) in the two SAG-associated mavirus and the absence in SV11_AB233_L11 mavirus of the gene 20 (152 amino acids, unknown function) of the endogenous mavirus genome (Fig. 4). Interestingly, we also retrieved an exon structure of one adjacent host gene of unknown function (gene 22, 177 amino acids) in the MAST-3A genome (Fig. 4). Although this putative host sequence is relatively short (~1kbp), we were able to observe a significant difference in its overall GC content compared with the mavirus sequence (60% *vs* 35%, respectively; Fig. 4).

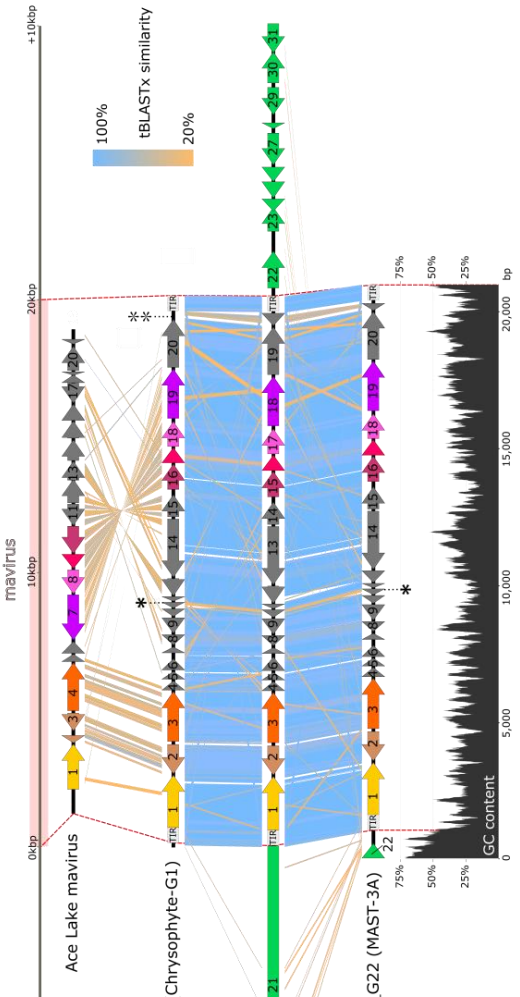


Figure 4. Comparison of the SAG-associated mavirus genomes and their closest known relatives. Linear genomic maps show synteny between the mavirus genomes found in chrysophyte-G1 and MAST-3A SAGs (SV11_AB233_L11 and SV11_AB240_G22, respectively) and their closest published relatives, endogenous mavirus and ALM (Zhou *et al.*, 2013). When present, TIRs and exon structures of putative adjacent host genes are displayed to highlight the putative integration of mavirus genomes within their respective host genomes. The main differences between the two SAG-associated mavirus and the endogenous mavirus are indicated with asterisks (*: presence of an extra coding gene [gene 11], *: absence of coding gene [gene 20]). Additionally, a GC content plot based on a 100 bp sliding window is shown for SV11_AB240_G22.

3.4. DISCUSSION

In this study, we used SCG to characterize potential virus-protist interactions. With the exception of *Pelagomonas calceolata*, we found evidence of virus associations in almost all studied protist cells. Indeed, the relatively high frequency of viral associations with protists cells (~57%), retrieved from SAG assemblies with low genome recovery (<10%), suggests that viral association levels are much higher. Same observations were previously made for prokaryotic cells in marine environments (Roux *et al.*, 2014; Labonté *et al.*, 2015; Munson-McGee *et al.*, 2018), implying that (nearly) all microbial cells are susceptible to be infected or to carry viruses. In the case of *Pelagomonas calceolata*, the lack of viral signals in the corresponding SAGs is probably due to the very low assembly coverage (0.5 ± 0.4%) rather than to the absence of any virus. However, we have not yet observed any significant correlation between genome completeness and the number of viral contigs among SAGs (Table 1). Similar findings were previously reported in bacterioplankton (Labonté *et al.*, 2015), suggesting that the probability to detect viruses among SAGs is independent of the retrieved host genome assembly. The variation in SAG genome coverage may depend on intrinsic properties of selected cells, their DNA integrity, as well as multiple displacement amplification (MDA) biases (Pinard *et al.*, 2006; Woyke *et al.*, 2009; Stepanauskas, 2012). Several methods have been developed to improve genome recovery of uncultured cells such as using partial SAG assemblies to recruit metagenome reads and/or contigs (Saw *et al.*, 2015), sorting multiple natural cells to perform a targeted metagenomic analysis (Cuvelier *et al.*, 2010; Vaulot *et al.*, 2012; Rinke *et al.*, 2013) or co-assembling short reads from multiple SAGs (Mangot *et al.*, 2017; Seeleuthner *et al.*, 2018). However, the application of these approaches to characterize virus–host interactions will miss intraspecific genetic variability of both actors. More recently, several new MDA-like methods, such as WGA-X (Stepanauskas *et al.*, 2017), TruePrime (Picher *et al.*, 2016) or REPLI-g (Ahsanuddin *et al.*, 2017) have been developed for improving the genome recovery from single environmental cells (bacterial, archaeal and protists) and viral particles with high GC-content

genomes. Compared with the conventional MDA, these amplification alternatives may provide a better genome recovery of microbial taxa, including some not amenable to standard MDA (Stepanauskas *et al.*, 2017).

Using VirSorter (Roux *et al.*, 2015) we were able to identify 64 unique viral contigs in 37 stramenopiles cells. We chose VirSorter over VirFinder (Ren *et al.*, 2017) because it has been shown that the later may misclassify eukaryotic sequences as viral. Some other approaches have been recently developed to retrieve viral signals from (meta-)genomic data, such as MARVEL (Amgarten *et al.*, 2018) and VirMiner (Zheng *et al.*, 2019), but they have been developed to detect viral genomes in prokaryotes. From the 64 viral contigs retrieved in the protist cells, the narrow host range of these viruses was remarkable given that >95% of the detected viral sequences (n=61) were specific to one stramenopile lineage and just a few were shared between lineages (n=3). This is contrary to previous findings on prokaryotic SAGs showing that nearly 50% of the detected viral types were found in more than 2 lineages (Munson-McGee *et al.*, 2018), suggesting that viruses infecting protists are likely more specialist than viruses infecting prokaryotes. Furthermore, while an important fraction (~54%) of cells with viral signals was associated to only one viral sequence, we also retrieved several putative co-infections among the remaining cells, with up to 7 unique (i.e., non-redundant) viral contigs in a single chrysophyte cell. Nonetheless, the risk of a putative accidental co-sorting of a free viral particle with a protist cell during the single-cell sorting process exists. To assess the risk of a possible "viral contamination", we estimated the frequency of such events based on previous estimates made on prokaryotic cells (Labonté *et al.*, 2015) by adapting the calculations to the cell size range of our studied stramenopile cells (2-3 µm). We obtained that the frequency of free environmental viral particles present in the cells' shade was less than 1 in 5,000. This reinforces the view that the viruses detected in our study were truly and directly associated to the analyzed protist cells. These associations may consist on i) lytic and/or temperate (i.e. nonlytic) viruses adsorbed to the cell membrane, ii) a temperate virus or a virophage

integrated into the host genome, iii) a virus replicating inside the cell, iv) a grazed prokaryote or protist carrying a temperate virus or with an active infection, or v) a predated free virus. A combination of these different scenarios probably explains the high number of viral sequences detected in the chrysophyte cells. Unfortunatelly, our current data set, including mostly fragments of viral sequences rather than complete viral genomes, does not allow us to decipher which mode of virus-host association prevail among the targeted protist cells.

Only 7 (~10%) viral contigs detected in protist cells were taxonomically assigned to known viruses (Table 2), which include some close hits to viruses belonging to the *Phycodnaviridae* family, known as a pathogen of marine eukaryotic algae (e.g., Brussaard, Short, Frederickson, & Suttle, 2004; Derelle et al., 2008), and others to bacteriophages and cyanophages. This suggests that a non-negligible part of the identified viral signals might come from putative infected (bacterial and/or picoeukaryotic) preys grazed by the stramenopiles. Indeed, the analyzed stramenopile lineages are mostly small free-living bacterivorous (Massana *et al.*, 2006; Piwosz *et al.*, 2013), with some groups (e.g., MAST-4) showing the ability to also eat picoalgae in grazing experiments (Massana *et al.*, 2009). Nevertheless, previous studies working with a subset of our SAGs (Mangot *et al.*, 2017; Seeleuthner *et al.*, 2018) have shown that genes from bacteria and photosynthetic eukaryotes only represent a very small fraction of the genome assemblies (< 0.3% of fragmented contigs (Mangot *et al.*, 2017)). A search for 16S rDNA genes in the SAGs where presumed bacteriophages were retrieved was unfruitful (data not shown), making difficult the association of these phages to putative grazed bacteria. It is also possible that some of the detected viral signals come from grazed viruses, since it is well-known that heterotrophic protists can graze on viruses (Gonzalez and Suttle, 1993; Fuhrman, 1999). However, another plausible explanation for the identification of bacteriophages as closest hits to some SAGs associated virus, is the overrepresentation of bacteriophage genomes compared to viruses in reference databases

(Klingenberg *et al.*, 2013), which is supported by the low sequence similarity between the viral signals and the bacteriophages sequences (Table 2). Although taking all together it is difficult to elucidate which virus-host associations prevail among the targeted protist cells, the geographic distribution of the viral signals supports the view that the detected virus-protists associations reflect in many cases true interactions, because viral signals coming from MAST-4A, MAST-4C and chrysophyte-H1 were ubiquitous (e.g. SV1, SV7 and SV28), while those viral signals coming from MAST-4E, MAST-3A, MAST-3F and chrysophyte-H2 were geographically constrained (e.g. SV12, SV32 and SV54) (Fig. 2), in agreement with the biogeography of these stramenopiles (Seeleuthner *et al.*, 2018).

Some of the taxonomically assigned viral contigs were affiliated to known virophages and, more particularly, in the case of SV11 to *Lavidaviridae* (Krupovic *et al.*, 2016). This virophage family, encompassing the two genera of *Mavirus* and *Sputnikvirus*, comprises obligate parasites of giant DNA viruses of the *Mimiviridae* family (Fischer and Hackl, 2016). Furthermore, virophages encode integrase genes, and provirophages have been reported in the nuclear genome of the marine alga *Bigelowiella natans* (Blanc *et al.*, 2015), and the protozoan *Cafeteria roenbergensis* (Fischer and Hackl, 2016). Provirophages putatively act as a host defense mechanism against giant viruses, in which some cells are sacrificed to protect their kin (Fischer and Suttle, 2011; Blanc *et al.*, 2015). In this study, we identified the presence of endogenous mavirus virophages in the assembly of five cells affiliated to chrysophyte-G1 and MAST-3A. These SAG-mavirus are highly similar to the *Cafeteriavirus*-dependent mavirus, a parasite of the giant *Cafeteria roenbergensis* virus (CroV) (Fischer *et al.*, 2010) integrated within the genome of *Cafeteria roenbergensis* (Fischer and Hackl, 2016). The presence of TIRs in the SV11_AB233_L11 and SV11_AB240_G22 virophages, as well as the exon structure of a putative adjacent host gene in the SV11_AB240_G22 sequence (Fig. 4), suggests the putative integration of the mavirus in the host genome. This is also confirmed by the lack of any CroV signal in our assemblies, whose presence is incompatible with a virophage in its

lysogenic stage (Fischer and Hackl, 2016). This finding constitutes the first report of the presence of a putative provirophage isolated from environmental samples using SCG. Only slight differences were observed between the different provirophage genomes, located notably at genomic regions of low conservation (gene 11 in the two SAG-associated mavirus). Little is known about the importance of mavirus provirophage in protist populations as its study is limited to few cases. It is somewhat surprising that the same mavirus virophage was found in three phylogenetically distant lineages (chrysophyte-G1, MAST-3A and *C. roenbergensis*), pointing to a global and important ecological role of virophages in protist populations. Mavirus host cell recognition is carried out through specific receptor interactions, while Sputnik entrance is done through phagocytosis of a composite of the virophage and the giant virus they parasitize (Duponchel and Fischer, 2019). Therefore, a possible explanation for finding mavirus in the different lineages, is that the capsid proteins are evolutionary conserved and have evolved independently of the giant virus infecting the host cell. On the contrary, although the host cells from Sputnik and Zamilon are phylogenetically closer, these virophages may have co-evolved with their corresponding giant virus. This hypothesis is supported by the finding that Sputnik can infect the groups A, B and C of the Mimiviridae group while Zamilon is unable to infect the group A (mimi- and mamavirus) (Gaia *et al.*, 2014). Virophages are repeatedly detected in genomic studies, with different gene content and abundance profiles, likely suggesting that they occupy different ecological niches (Yau *et al.*, 2011; Desnues and Raoult, 2012; Roux *et al.*, 2017). Although the role of virophages in protist populations is still enigmatic, they may play a role in regulating the giant virus population dynamics and virus-host interactions, influencing the ecosystem function and probably the whole microbial food web in aquatic environments (Desnues and Raoult, 2012). Our findings provide new insights into the potential importance of mavirus in the ecology of marine protists, and reinforce the need for more studies to elucidate the role of these fascinating viruses in the environment.

In summary, this work shows the benefits of single-cell genomics to increase our understanding of virus-host associations in natural protist communities. Although our knowledge of the marine viral diversity is constantly expanding since the development of metagenomics (Mizuno *et al.*, 2013; Paez-Espino *et al.*, 2016; Coutinho *et al.*, 2017), it has been estimated that the majority (63-93%) of viral sequences in marine metagenomes are not represented in public databases (Hurwitz and Sullivan, 2013), emphasizing the need for further isolation, characterization and sequencing of specific marine viruses (Middelboe and Brussaard, 2017). A minute fraction of protist viruses is annotated to date (~100 sequenced genomes, ~0.6% of all viral genomes) in NCBI Genome database (July 2018), explaining the majority of unassigned viral sequences in our study. Thus, in addition to the ever-increasing knowledge on viral diversity by metagenomic approaches, the incorporation of SAG analysis will allow the specific matching of viruses and their hosts as well as to determine the host range of individual viruses without cultivation. Our findings suggest that protist cells are susceptible to interact with predominantly specialist viruses and hint to the potential importance of provirophages in protist populations.

REFERENCES

Ahsanuddin, S., Afshinnekoo, E., Gandara, J., Hakyemezoğlu, M., Bezdan, D., Minot, S., et al. (2017) Assessment of REPLI-g multiple displacement whole genome amplification (WGA) techniques for metagenomic applications. *J. Biomol. Tech.* **28**: 46–55.

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., et al. (2017) Viral to metazoan marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci. Data* **4**: 170093.

Allers, E., Moraru, C., Duhaime, M.B., Beneze, E., Solonenko, N., Barrero-Canosa, J., et al. (2013) Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environ. Microbiol.* **15**: 2306–2318.

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

Amgarten, D., Braga, L.P.P., da Silva, A.M., and Setubal, J.C. (2018) MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Front. Genet.* **9**: 304.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol. Ecol.* **77**: 120–133.

Baran, N., Goldin, S., Maidanik, I., and Lindell, D. (2018) Quantification of diverse virus populations in the environment using the polony method. *Nat. Microbiol.* **3**: 62–72.

Berg Miller, M.E., Yeoman, C.J., Chia, N., Tringe, S.G., Angly, F.E., Edwards, R.A., et al. (2012) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ. Microbiol.* **14**: 207–227.

Bhattacharya, D., Price, D.C., Yoon, H.S., Yang, E.C., Poulton, N.J., Andersen, R.A., and Das, S.P. (2012) Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. *Sci. Rep.* **2**: 356.

Blanc, G., Gallot-Lavallée, L., and Maumus, F. (2015) Provirophages in the *Bigelowiella* genome bear testimony to past encounters with giant viruses. *Proc. Natl. Acad. Sci.* **112**: E5318–E5326.

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.

Breitbart, M. (2012) Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* **4**: 425–448.

Breitbart, M., Bonnain, C., Malki, K., and Sawaya, N.A. (2018) Phage puppet

masters of the marine microbial realm. *Nat. Microbiol.* **3**: 754–766.

Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., et al. (2015) Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498–1261498.

Brum, J.R. and Sullivan, M.B. (2015) Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**: 147–159.

Brussaard, C.P.D., Short, S.M., Frederickson, C.M., and Suttle, C.A. (2004) Isolation and phylogenetic analysis of novel viruses infecting the phytoplankton *Phaeocystis globosa* (Prymnesiophyceae). *Appl. Environ. Microbiol.* **70**: 3700–3705.

Brussaard, C.P.D., Wilhelm, S.W., Thingstad, F., Weinbauer, M.G., Bratbak, G., Heldal, M., et al. (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J.* **2**: 575–578.

Brüssow, H. and Hendrix, R.W. (2002) Phage genomics: small is beautiful. *Cell* **108**: 13–16.

del Campo, J. and Massana, R. (2011) Emerging diversity within Chrysophytes, Choanoflagellates and Bicosoecids based on molecular surveys. *Protist* **162**: 435–448.

Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Chow, C.E.T., Winget, D.M., White, R.A., Hallam, S.J., and Suttle, C.A. (2015) Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front. Microbiol.* **6**: 1–15.

Coutinho, F.H., Silveira, C.B., Gregoracci, G.B., Thompson, C.C., Edwards, R.A., Brussaard, C.P.D., et al. (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat. Commun.* **8**: 15955.

Cuvelier, M.L., Allen, A.E., Monier, A., McCrow, J.P., Messie, M., Tringe, S.G., et al. (2010) Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci.* **107**: 14679–14684.

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J.A., Middelburg, J.J., Noble, R.T., and Suttle, C.A. (2011) Marine viruses and global climate change. *FEMS Microbiol. Rev.* **35**: 993–1034.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**: 1164–1165.

Deng, L., Gregory, A., Yilmaz, S., Poulos, B.T., Hugenholtz, P., and Sullivan, M.B. (2012) Contrasting life strategies of viruses that infect photo- and

heterotrophic bacteria, as revealed by viral tagging. *MBio* **3**: e00373-12.

Deng, L., Ignacio-Espinoza, J.C., Gregory, A.C., Poulos, B.T., Weitz, J.S., Hugenholtz, P., and Sullivan, M.B. (2014) Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* **513**: 242–245.

Derelle, E., Ferraz, C., Escande, M.-L., Eychenié, S., Cooke, R., Piganeau, G., et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS One* **3**: e2250.

Desnues, C. and Raoult, D. (2012) Virophages question the existence of satellites. *Nat. Rev. Microbiol.* **10**: 234–234.

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., et al. (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc. Natl. Acad. Sci.* **109**: 18078–18083.

Devisetty, U.K., Kennedy, K., Sarando, P., Merchant, N., and Lyons, E. (2016) Bringing your tools to CyVerse Discovery Environment using Docker. *F1000Research* **5**: 1442.

Duponchel, S. and Fischer, M.G. (2019) Viva lavidaviruses! Five features of virophages that parasitize giant DNA viruses. *PLOS Pathog.* **15**: e1007592.

Fischer, M.G., Allen, M.J., Wilson, W.H., and Suttle, C.A. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl. Acad. Sci.* **107**: 19508–19513.

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

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

Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541–548.

Gaia, M., Benamar, S., Boughalmi, M., Pagnier, I., Croce, O., Colson, P., et al. (2014) Zamilon, a novel virophage with Mimiviridae host specificity. *PLoS One* **9**: e94923.

Gaia, M., Pagnier, I., Campocasso, A., Fournous, G., Raoult, D., and La Scola, B. (2013) Broad spectrum of Mimiviridae virophage allows its isolation using a Mimivirus reporter. *PLoS One* **8**: e61912.

Gong, C., Zhang, W., Zhou, X., Wang, H., Sun, G., Xiao, J., et al. (2016) Novel virophages discovered in a freshwater lake in China. *Front. Microbiol.* **7**: 5.

Gonzalez, J.M. and Suttle, C.A. (1993) Grazing by marine nanoflagellates on viruses and virus-sized particles: ingestion and digestion.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–1075.

Heywood, J.L., Sieracki, M.E., Bellows, W., Poulton, N.J., and Stepanauskas, R.

(2011) Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J.* **5**: 674–684.

Hurwitz, B.L. and Sullivan, M.B. (2013) The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**: e57355.

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.

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

Karsenti, E., Acinas, S.G., Bork, P., Bowler, C., De Vargas, C., Raes, J., et al. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol.* **9**: e1001177.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**: 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Klingenberg, H., Aßhauer, K.P., Lingner, T., and Meinicke, P. (2013) Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics* **29**: 973–980.

Krabberød, A., Bjorbækmo, M., Shalchian-Tabrizi, K., and Logares, R. (2017) Exploring the oceanic microeukaryotic interactome with metaomics approaches. *Aquat. Microb. Ecol.* **79**: 1–12.

Krupovic, M., Kuhn, J.H., and Fischer, M.G. (2016) A classification system for virophages and satellite viruses. *Arch. Virol.* **161**: 233–247.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* **9**: 2386–2399.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Lin, Y.-C.C., Campbell, T., Chung, C.-C.C., Gong, G.-C.C., Chiang, K.-P.P., and Worden, A.Z. (2012) Distribution patterns and phylogeny of marine stramenopiles in the North Pacific Ocean. *Appl. Environ. Microbiol.* **78**: 3387–3399.

Logares, R., Audic, S., Santini, S., Pernice, M.C., de Vargas, C., and Massana, R. (2012) Diversity patterns and activity of uncultured marine heterotrophic

flagellates unveiled with pyrosequencing. *ISME J.* **6**: 1823–1833.

Mangot, J.-F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S., et al. (2017) Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **7**: 41498.

Massana, R. (2011) Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**: 91–110.

Massana, R., del Campo, J., Sieracki, M.E., Audic, S., and Logares, R. (2014) Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8**: 854–866.

Massana, R., Terrado, R., Forn, I., Lovejoy, C., and Pedrós-Alió, C. (2006) Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environ. Microbiol.* **8**: 1515–1522.

Massana, R., Unrein, F., Rodríguez-Martínez, R., Forn, I., Lefort, T., Pinhassi, J., and Not, F. (2009) Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J.* **3**: 588–596.

Middelboe, M. and Brussaard, C. (2017) Marine viruses: key players in marine ecosystems. *Viruses* **9**: 302.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016) Linking virus genomes with host taxonomy. *Viruses* **8**: 66.

Mizuno, C.M., Rodriguez-Valera, F., Kimes, N.E., and Ghai, R. (2013) Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**: e1003987.

Munn, C.B. (2006) Viruses as pathogens of marine organisms—from bacteria to whales. *J. Mar. Biol. Assoc. UK* **86**: 453–467.

Munson-McGee, J.H., Peng, S., Dewerff, S., Stepanauskas, R., Whitaker, R.J., Weitz, J.S., and Young, M.J. (2018) A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *ISME J.* **12**: 1706–1714.

Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., et al. (2017) Environmental viral genomes shed new light on virus-host interactions in the ocean. *mSphere* **2**: e00359-16.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A., et al. (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**: 714–737.

Oh, S., Yoo, D., and Liu, W.-T. (2016) Metagenomics reveals a novel virophage population in a tibetan mountain lake. *Microbes Environ.* **31**: 173–177.

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.

Paradis, E., Claude, J., and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

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.

Picher, Á.J., Budeus, B., Wafzig, O., Krüger, C., García-Gómez, S., Martínez-Jiménez, M.I., et al. (2016) TruePrime is a novel method for whole-genome amplification from single cells based on TthPrimPol. *Nat. Commun.* **7**: 13296.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**: 216.

Piwosz, K., Wiktor, J.M., Niemi, A., Tatarek, A., and Michel, C. (2013) Mesoscale distribution and functional diversity of picoeukaryotes in the first-year sea ice of the Canadian Arctic. *ISME J.* **7**: 1461–1471.

R Development Core Team (2016) R: a language and environment for statistical computing.

Rappé, M.S. and Giovannoni, S.J. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**: 369–394.

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

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

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.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689–693.

Roux, S., Chan, L.-K., Egan, R., Malmstrom, R.R., McMahon, K.D., and Sullivan, M.B. (2017) Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat. Commun.* **8**: 858.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**: e985.

Roux, S., Hawley, A.K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., et al. (2014) Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *Elife* **3**: e03125.

Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S., et al.

(2015) Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep.* **4**: 4780.

Saw, J.H., Spang, A., Zaremba-Niedzwiedzka, K., Juzokaite, L., Dodsworth, J.A., Murugapiran, S.K., et al. (2015) Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**: 20140328.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., et al. (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* **455**: 100–104.

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., et al. (2018) Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun.* **9**: 310.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**: 613–620.

Stepanauskas, R., Fergusson, E.A., Brown, J., Poulton, N.J., Tupper, B., Labonté, J.M., et al. (2017) Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* **8**: 84.

Sullivan, M.J., Petty, N.K., and Beatson, S.A. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics* **27**: 1009–1010.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., et al. (2015) Structure and function of the global ocean microbiome. *Science* **348**: 1261359–1261359.

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

Swan, B.K., Tupper, B., Sczyrba, A., Lauro, F.M., Martinez-Garcia, M., Gonzalez, J.M., et al. (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci.* **110**: 11463–11468.

Tadmor, A.D., Ottesen, E.A., Leadbetter, J.R., and Phillips, R. (2011) Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**: 58–62.

Troell, K., Hallström, B., Divne, A.-M., Alsmark, C., Arrighi, R., Huss, M., et al. (2016) *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics* **17**: 471.

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.-M., et al. (2016) Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Sci. Rep.* **6**: 37900.

Vaulot, D., Lepère, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., et al. (2012) Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* **7**: e39648.

Weitz, J. and Wilhelm, S. (2012) Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* **4**: 17.

Woyke, T., Xie, G., Copeland, A., González, J.M., Han, C., Kiss, H., et al. (2009) Assembling the marine metagenome, one cell at a time. *PLoS One* **4**: e5299.

Yau, S., Lauro, F.M., DeMaere, M.Z., Brown, M. V., Thomas, T., Raftery, M.J., et al. (2011) Virophage control of antarctic algal host-virus dynamics. *Proc. Natl. Acad. Sci.* **108**: 6163–6168.

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., et al. (2011) Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**: 714–717.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017) ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**: 28–36.

Zheng, T., Li, J., Ni, Y., Kang, K., Misiakou, M.-A., Imamovic, L., et al. (2019) Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* **7**: 42.

Zhou, J., Sun, D., Childers, A., McDermott, T.R., Wang, Y., and Liles, M.R. (2015) Three novel virophage genomes discovered from Yellowstone Lake metagenomes. *J. Virol.* **89**: 1278–1285.

Zhou, J., Zhang, W., Yan, S., Xiao, J., Zhang, Y., Li, B., et al. (2013) Diversity of virophages in metagenomic data sets. *J. Virol.* **87**: 4225–4236.

SUPPLEMENTARY INFORMATION

Supplementary tables

Table S1. General characteristics of the analysed SAGs.

Group	Name	SAG ID	Trophic mode	Station	Depth	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)	Number of viral contigs
Chrysophyceae												
	Chrysophyte-G1	AB233_O05	Phototrophy	41	Surface	4.7	7.9	3,130	8.4	40.1	4.7	4
		AB233_P23				5.5	17.1	6,520	23.7	40.5	4.9	1
		AB233_L11				5.7	6.9	2,771	8.8	40.1	5.3	1
		AB233_D06				5.9	5.6	1,966	5.6	40.0	5.7	2
	Chrysophyte-H1[†]	AA538_K15	Heterotrophy	23	DCM	3.1	3.1	1,574	4.2	44.3	5.6	1
		AA538_D22				3.1	1.7	1,015	3.3	43.9	3.2	1
		AA538_I04				3.8	1.9	646	1.9	44.3	10.2	1
		AA538_C03				4.1	4.1	1,459	6.0	45.6	7.1	0
		AA538_K19				4.3	5.9	1,726	8.8	45.9	12.8	9[§]
		AA538_G15				4.4	6.0	1,770	7.4	45.9	11.6	5
		AA538_D14				4.5	2.6	968	4.7	45.3	7.1	3
		AA538_J08				4.9	7.1	2,239	12.6	45.7	11.3	4

Table S1. Continuation.

Group	Name	SAG ID	Trophic mode	Station	Depth	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)	Number of viral contigs
	Chrysophyte-H2^{+}	AA538_A16	Heterotrophy	23	DCM	3.1	5.1	2,196	3.3	48.7	4.4	6
		AA538_J21				4.6	6.3	2,841	5.1	48.7	4.0	5
		AA538_P21				5	1.6	746	1.4	45.8	3.9	1
Dictyo												
	unc. clade	AB206_B02	Phototrophy	47	Surface	4.4	0.8	484	0.0	43.3	2.6	0
		AB206_F09				4.6	3.1	1,420	0.9	47.8	4.8	0
		AB198_G07		48		4.6	3.1	1,531	0.9	47.3	3.7	2
		AB198_K18				4.7	7.5	2,792	2.3	48.9	5.7	1
MAST-3												
	MAST-3A	AB240_P16	Heterotrophy	41	Surface	4.7	8.8	2,580	12.1	42.7	9.6	1
		AB241_O20				4.9	4.6	1,688	7.9	42,0	7.1	1
		AB241_L22				5	8.4	2,528	15.8	42.6	8.4	2
		AB240_G22				5.9	8.0	2,292	9.8	42.6	9.3	2
	MAST-3F	AA538_B10	Heterotrophy	23	DCM	4.6	5.4	1,869	4.7	34.3	7.5	1
		AA538_E07				6.1	16.7	5,283	18.6	33.9	7.9	0
MAST-4												
	MAST-4A	AA538_E19	Heterotrophy	23	DCM	2.4	9.7	3,316	13.0	32.5	7.4	0
		AA538_C11				2.7	17.5	5,692	26.0	32.6	8.6	0

Table S1. Continuation.

Group	Name	SAG ID	Trophic mode	Station	Depth	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)	Number of viral contigs
	MAST-4A (contd)	AA538_J18				3.2	8.3	3,590	6.5	32.5	5.6	0
		AA538_K07				4	2.0	939	2.3	36.5	4.0	1
		AA538_G20‡				4.6	10.2	3,121	10.7	32.7	10.7	2
		AA538_G04				4.7	9.3	3,014	9.8	32.5	8.5	1
		AA538_E21				5.7	22.2	6,040	30.2	32.7	11.5	0
		AA538_F10				6	10.6	3,193	11.6	32.5	9.5	2
		AA538_E15				6.4	10.4	3,026	15.3	33,0	10.1	3
		AA538_G20_bis‡				6.8	11.5	3,709	11.6	32.8	8.6	1
		AA538_M19				6.9	10.0	2,994	8.8	33.2	10.8	0
		AA538_N22				8.4	8.3	2,452	10.2	32.6	9.2	0
		AB537_A17		41		4	8.0	2,515	14.9	32.6	10.2	1
		AB537_K04				3.5	4.8	1,133	6.0	32.5	15.4	1
	MAST-4C	AB536_M21	Heterotrophy	41	DCM	4.7	5.2	1,545	7.0	40.2	13.0	0
		AB536_J08				5.1	8.2	2,485	11.6	40.5	12.9	1
		AB536_F22				5.5	10.6	2,722	12.1	40.5	15.4	0
		AB536_E17				6.1	9.4	2,806	16.3	40.1	15.1	0
	MAST-4E	AA538_F08	Heterotrophy	23	DCM	4	7.2	2,048	9.8	44.5	9.2	0
		AA538_M11				4.2	2.6	854	3.7	42.5	7.4	0
		AA538_L23				4.4	3.3	1,176	3.3	43.5	6.9	1

Table S1. Continuation.

Group	Name	SAG ID	Trophic mode	Station	Depth	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)	Number of viral contigs
	MAST-4E (contd)	AA538_A02				4.5	7.9	2,190	7.0	44.2	10.2	3
		AA538_A03				4.5	8.4	2,085	13.5	44.1	12.1	0
		AA538_C05				4.6	7.6	2,194	11.2	44.5	10.6	0
		AA538_J09				4.7	7.8	2,013	12.6	44.3	10.5	0
		AA538_N16				4.9	5.3	1,845	5.1	43.6	6.8	1
		AA538_A11				6.8	10.6	2,948	11.6	44.4	10.0	0
MAST-7												
	MAST-7A	AA538_I11	Heterotrophy	23	DCM	3.9	1.0	561	0.9	44.2	3.1	1
		AA538_B21				4.2	6.3	2,262	5.1	48.5	7.0	2
		AA538_D10				6.7	2.8	1,018	2.8	36.5	5.8	1
		AA538_M21				7.2	10.1	4,122	3.7	42.7	4.9	3
		AB536_L18		41		5.1	6.8	2,037	3.3	46.2	10.1	0
		AB538_M04				5.7	6.6	2,012	7.0	49.9	11.4	0
Pelago												
	Pelagomonas calceolata	AA534_B13	Phototrophy	23	DCM	4.9	0.4	223	0.0	45.9	4.8	0
		AA534_D23				5.2	0.3	116	0.0	42.5	34.6	0
		AA534_I02				5.7	1.6	587	0.5	56.9	12.3	0
		AA534_N20				6.3	2.2	467	2.8	58.9	20.2	0
		AA534_P10				6.2	0.4	149	0.0	42	7.6	0

Table S1. Continuation.

Group	Name	SAG ID	Trophic mode	Station	Depth	Sequencing depth (Gbp)	Assembly size (Mbp)	Total number of contigs	BUSCO completeness (%)	GC content (%)	N50 (kbp)	Number of viral contigs
	Pelagomonas calceolata (contd)	AA534_F21				5.5	0.3	107	0.0	39.4	5.7	0
		AA534_I03				5.7	0.6	247	0.0	47	6.0	0

Abbreviations: SAG, Single Amplified Genome; BUSCO, Benchmarking Universal Single-Copy Orthologs); N50, length of the shortest contig from the minimal set of contig representing 50% of the assembly size; DCM, Deep Chlorophyll Maximum; unc., uncultured; Dictyo, Dictyochophyceae; MAST, Marine Stramenopiles; contd, continued; Pelago, Pelagophyceae.

[†]The 18S rRNA genes of these chrysophytes-H SAGs clustered into two distinct lineages (clades -H1 and -H2).

[‡]SAG sequenced by two different sequencing centers. The two sequencing replicates (AA538_G20 and AA538_G20_bis) were kept for further analysis.

[§]One viral sequence was present in triplicate in this SAG.

Table S2. Accession codes of the analysed SAGs.

Group	Name	SAG ID	Accession code	Scientific name
Chrysophyceae				
	Chrysophyte-G1	AB233_D06	ERR3417514	Chrysophyceae sp.
		AB233_L11	ERR3417515	TOSAG41-3
		AB233_O05	ERR3417516	
		AB233_P23	ERR3417517	
	Chrysophyte-H1†	AA538_K15	ERR1189849	Chrysophyceae sp.
		AA538_D22	ERR1189855	TOSAG23-4
		AA538_C03	ERR1198956	
		AA538_D14	ERR1198934	
		AA538_G15	ERR1198924	
		AA538_I04	ERR1198937	
		AA538_J08	ERR1198933	
		AA538_K19	ERR1198951	
	Chrysophyte-H2†	AA538_A16	ERR1198944	Chrysophyceae sp.
		AA538_J21	ERR1198935	TOSAG23-5
		AA538_P21	ERR1198929	
Dictyochophyceae				
	unc. dictyochophyte	AB206_B02	ERR3438858	Dictyochophyceae sp.
		AB206_F09	ERR3417524	TOSAG47-1
		AB198_G07	ERR3417522	Dictyochophyceae sp.
		AB198_K18	ERR3417523	TOSAG48-1
MAST-3				
	MAST-3A	AB241_L22	ERR1198953	Stramenopiles sp.
		AB241_O20	ERR1198930	TOSAG41-2
		AB240_P16	ERR1198943	
		AB240_G22	ERR1198931	
	MAST-3F	AA538_B10	ERR1189848	Stramenopiles sp.
		AA538_E07	ERR1189852	TOSAG23-6
MAST-4				
	MAST-4A	AA538_M19	ERR1198936	Stramenopiles sp.
		AA538_F10	ERR1198948	TOSAG23-1
		AA538_G04	ERR1198954	
		AA538_K07	ERR1198938	
		AA538_G20‡	ERR1198925	
		AA538_N22	ERR1198949	
		AA538_C11	ERR1138643	Stramenopiles sp.
		AA538_E15	ERR1138644	TOSAG23-2
		AB537-A17	ERR1138645	
		AA538_E21	ERR1138646	
		AA538_E19	ERR1744380	Stramenopiles sp.
		AA538_J18	ERR1744377	TOSAG23-1
		AA538_G20_bis‡	ERR1744378	
		AB537_K04	ERR1744379	
	MAST-4C	AB536_E17	ERR1198955	Stramenopiles sp.
		AB536_F22	ERR1198945	TOSAG41-1
		AB536_J08	ERR1198926	

Table S2. Continuation.

Group	Name	SAG ID	Accession code	Scientific name
	MAST-4C (contd)	AB536_M21	ERR1198940	
	MAST-4E	AA538_A03	ERR1189846	Stramenopiles sp.
		AA538_C05	ERR1189854	TOSAG23-3
		AA538_F08	ERR1189844	
		AA538_J09	ERR1189847	
		AA538_A11	ERR1198928	
		AA538_L23	ERR1198946	
		AA538_M11	ERR1198927	
		AA538_N16	ERR1198941	
		AA538_A02	ERR1198950	
MAST-7				
	MAST-7A	AA538_D10	ERR3417518	Stramenopiles sp. MAST-7
		AA538_M21	ERR3417519	TOSAG23-8
		AA538_I11	ERR3417521	Stramenopiles sp. MAST-7
		AA538_B21	ERR3417520	TOSAG23-7
		AB536_L18	ERR3417525	Stramenopiles sp. MAST-7
		AB538_M04	ERR3417526	TOSAG41-4
Pelagophyceae				
	Pelagomonas calceolata	AA534_B13	ERR3438851	*Pelagomonas calceolata*
		AA534_D23	ERR3438852	TOSAG23-9
		AA534_I02	ERR3438854	
		AA534_N20	ERR3438856	
		AA534_P10	ERR3438857	
		AA534_F21	ERR3438853	
		AA534_I03	ERR3438855	

Abbreviations: SAG, Single Amplified Genome; MAST, Marine Stramenopiles; contd, continued.

† The 18S rRNA genes of these chrysophytes-H SAGs clustered into two distinct lineages (clades -H1 and -H2).

‡ SAG sequenced by two different sequencing centers. The two sequencing replicates (AA538_G20 and AA538_G20_bis) were kept for further analysis.

Table S3. Description of the locations visited during the *Tara* Oceans expedition where the cells for the SCGs were collected.

St	Depth	Date	Location	Latitude	Longitude	Sample depth (m)	Temp (°C)	Salinity (psu)
23	DCM	18/11/2009	Adriatic Sea	42° 11' 23.5" N	17° 43' 0.12" E	55.2	17.3	38.2
41	Surface	30/03/2010	Indian Ocean	14° 35' 43.44" N	69° 58' 51.6" E	3	29.1	36
41	DCM	30/03/2010	Indian Ocean	14° 35' 43.44" N	69° 58' 51.6" E	59.3	27.2	36.5
47	Surface	16/04/2010	Indian Ocean	-2° 2' 47.51" N	72° 9' 24.48" E	3	30.2	34.9
48	Surface	19/04/2010	Indian Ocean	-9° 24' 10.62" N	66° 22' 4.94" E	3	29.8	34.2

Abbreviations: St, Station; Temp, Temperature;

Table S4. Predicted gene function of the 64 SAG-associated viral sequences. Best BLAST hits to eukaryotic viruses (*****), virophages (******) and bacteriophages (**§**) are marked.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV1	1	Bifunctional DNA primase/helicase	2e-149	79.1	Candidatus Bathyarchaeota archaeon (RLI52043)	
	2	Hypothetical protein	1e-4	36.2	unc. virus (ASF00535)	*
	3	Hypothetical protein DWQ49_06995	2e-24	65.2	Bacteroidetes bacterium (REK60078)	
	4	Thymidylate synthase	8e-123	62.3	Bacteroidetes bacterium (REK60077)	
	5	Hypothetical protein DRO61_12470	7e-75	45.2	Candidatus Bathyarchaeota archaeon (RLI44166)	
	6	DUF1064 domain-containing protein	7e-68	67.8	Bacteroidetes bacterium (REK60079)	
	7	-	-	-	-	
	8	Hypothetical protein CBB75_13870	0	73.3	bacterium TMED15 (OUT57717)	
	9	-	-	-	-	
	10	Hypothetical protein	7e-27	32.2	Flavobacterium psychrophilum (WP_094138647)	
	11	Hypothetical protein CBD27_11755, partial	1e-115	38.9	Rhodospirillaceae bacterium TMED167 (OUW23939)	
	12	-	-	-	-	
	13	Hypothetical protein DWQ21_06100	7e-60	60.3	Bacteroidetes bacterium (REJ62561)	
	14	Hypothetical protein CBD16_04705	2e-34	39.9	Betaproteobacteria bacterium TMED156 (OUW01968)	
	15	Hypothetical protein EHM12_12210, partial	4e-172	51.8	Dehalococcoidia bacterium (RPJ55435)	
	16	Hypothetical protein DWQ21_06115	0	37.5	Bacteroidetes bacterium (REJ62564)	
	17	Hypothetical protein CBC27_06610	4e-161	59.9	Opitutae bacterium TMED67 (OUU71870)	
	18	Putative structural protein	5e-59	39	unc. virus (ASF00183)	*
	19	Hypothetical protein DRN17_07930, partial	3e-130	67.3	Thermoplasmata archaeon (RLF42319)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV1	20	Hypothetical protein CBC27_06595	0	74.4	Opitutae bacterium TMED67 (OUU71867)	
(contd)	21	Hypothetical protein DRO91_09215, partial	0	90.2	Candidatus Heimdallarchaeota archaeon (RLI68315)	
	22	Hypothetical protein DRO91_09210, partial	0	91	Candidatus Heimdallarchaeota archaeon (RLI68314)	
	23	Hypothetical protein DRI84_07185	0	75.3	Bacteroidetes bacterium (RLD65262)	
	24	Hypothetical protein DRO61_08960	4e-44	82.2	Candidatus Bathyarchaeota archaeon (RLI46659)	
	25	Hypothetical protein DRO61_08955, partial	4e-104	89	Candidatus Bathyarchaeota archaeon (RLI46658)	
	26	Hypothetical protein B7C24_09080	2e-12	45.7	Bacteroidetes bacterium 4572_77 (OYT16199)	
	27	-	-	-	-	
	28	-	-	-	-	
	29	Hypothetical protein P12024L_15	1e-17	38.2	unc. marine virus (AKH47409)	*
	30	Hypothetical protein CBC27_06660	4e-29	69.5	Opitutae bacterium TMED67 (OUU71879)	
	31	Hypothetical protein CBC48_13745	6e-68	76.2	unc. bacteria (OUV27981)	
	32	Hypothetical protein CBC27_08435	1e-10	60	unc. Opitutae (OUU70618)	
	33	-	-	-	-	
	34	Hypothetical protein CBC30_00115	2e-60	56.8	Chloroflexi bacterium TMED70 (OUU78042)	
	35	Co-chaperonin groes	5e-37	73.3	unc. virus (ASN63467)	*
	36	Chaperonin groel	0	85.9	unc. virus (AQM32683)	*
	37	Hypothetical protein CBC27_08495	6e-39	65.4	unc. Opitutae (OUU70630)	
	38	Hypothetical protein CBC27_08485	9e-97	37.7	unc. Opitutae (OUU70628)	
	39	Hypothetical protein DRO61_12040	3e-12	41.1	Candidatus Bathyarchaeota (RLI44538)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV1	40	Hypothetical protein	5e-32	55	unc. virus (AMQ66422)	*
(contd)	41	Hypothetical protein	4e-24	33.4	*Chryseobacterium* sp. (WP_101240788)	
	42	Hypothetical protein DWQ21_06245	2e-09	53.3	unc. Bacteroidetes (REJ62589)	
	43	Hypothetical protein CBC27_03795	5e-06	53.8	unc. Opitutae (OUU73313)	
	44	-	-	-	-	
SV2	1	Phage-related tail fiber protein	1e-132	67.8	unc. Mediterranean phage (BAR30060)	§
	2	Hypothetical protein C4K49_10520 (partial)	3e-77	35	candidatus Thorarchaeota (RDE12107)	
	3	Mannosidase-related	2e-80	62.8	unc. Mediterranean phage (BAR30144)	§
	4	Hypothetical protein	5e-15	44.1	unc. Mediterranean phage (BAR14092)	§
	5	Putative carbohydrate binding domain containing protein	1e-101	52.2	unc. Mediterranean phage (BAQ93964)	§
	6	Hypothetical protein meddcm-OCT-S15-C1-cds21	7e-09	55	unc. Mediterranean phage (AFX83761)	§
	7	Hypothetical protein	3e-34	72	unc. Mediterranean phage (BAR31749)	§
	8	Hypothetical protein	1e-08	46.3	unc. Mediterranean phage (BAR30991)	§
	9	Hypothetical protein	9e-10	44	unc. phage (ADD94579)	§
	10	-	-	-	-	
	11	Hypothetical protein CPTG_00059	1e-29	39.8	Cyanophage (AGH56352)	§
	12	Gp165	4e-54	50.6	unc. Mediterranean phage (BAR38762)	§
	13	Hypothetical protein	3e-40	50.8	Cyanophage (YP_007006163)	§
	14	Hypothetical protein RW270310_155	6e-54	60.7	Cyanophage (AOO17574)	
	15	Hypothetical protein CBC18_06465	5e-33	51.4	unc. Deltaproteobacteria (OUU32209)	
	16	Hypothetical protein	6e-06	43.1	unc. Mediterranean phage (BAR21101)	§

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV2	17	Hypothetical protein meddcm-OCT-S13-C2-cds7	1e-17	66.1	unc. Mediterranean phage (AFX83624)	§
(contd)	18	Hypothetical protein	2e-05	46.3	unc. Mediterranean phage (BAR36173)	§
	19	-	-	-	-	
	20	-	-	-	-	
	21	Hypothetical protein	4e-07	39.7	unc. Mediterranean phage (BAR14416)	§
	22	Hypothetical protein	3e-05	45.3	unc. Mediterranean phage (ANS05219)	§
	23	-	-	-	-	
	24	Hypothetical protein BL107_10711	1e-07	36.5	unc. phage (ADD95153)	§
	25	Hypothetical protein	2e-04	41	unc. Mediterranean phage (BAQ93266)	§
	26	-	-	-	-	
	27	Hypothetical protein	2e-30	58.6	unc. Mediterranean phage (ANS05217)	§
	28	-	-	-	-	
	29	Rad52 recombinase	4e-75	51.2	unc. Mediterranean phage (BAQ92306)	§
	30	Holliday junction resolvase rusa like	1e-50	58.4	unc. Mediterranean phage (BAQ92307)	§
	31	Parb-like nuclease domain containing protein	2e-58	36.1	unc. Mediterranean phage (BAR14373)	§
	32	DNA replication factor Dna2-like nuclease	4e-80	57.5	unc. Mediterranean phage (ANS05214)	§
	33	-	-	-	-	
	34	Hypothetical protein	2e-29	60.2	unc. Mediterranean phage (ANS05213)	§
	35	Hypothetical protein	3e-05	35	unc. Mediterranean phage (BAQ92312)	§
	36	Hypothetical protein	9e-23	39.1	unc. Mediterranean phage (ANS05212)	§
	37	-	-	-	-	
	38	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV2	39	-	-	-	-	
(contd)	40	-	-	-	-	
	41	-	-	-	-	
	42	-	-	-	-	
	43	Hypothetical protein PRRG_00004	6e-18	43.9	Prochlorococcus phage (AGF91515)	§
	44	-	-	-	-	
	45	Hypothetical protein DWQ28_02650	7e-11	38.5	unc. Proteobacteria (REJ71121)	
	46	-	-	-	-	
	47	-	-	-	-	
	48	-	-	-	-	
SV3	1	-	-	-	-	
	2	-	-	-	-	
	3	-	-	-	-	
	4	-	-	-	-	
	5	Hypothetical protein YSLV5_ORF11	5e-21	29.6	YSLV5 (YP_009177794)	**
	6	Hypothetical protein B4U80_15061	2e-23	44.4	*Leptotrombidium deliense* (RWS13789)	
	7	-	-	-	-	
	8	Hypothetical protein TRIADDRAFT_62442	7e-33	27.9	*Trichoplax adhaerens* (XP_002118406)	
	9	-	-	-	-	
	10	Hypothetical protein trisph2_011690	2e-23	30.2	*Trichoplax* sp. (RDD36227)	
	11	-	-	-	-	
	12	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV3 (contd)	13	-	-	-	-	
	14	Hypothetical protein B4U80_13961	6e-35	35.9	*Leptotrombidium deliense* (RWS23085)	
	15	Putative DNA primase/polymerase	8e-35	28.5	OLV (ADX05784)	**
	16	-	-	-	-	
	17	-	-	-	-	
	18	-	-	-	-	
	19	Trna (adenine-N(6)-)-methyltransferase	3e-11	34.5	*Eubacterium eligens* (WP_118370243)	
	20	Sugar-phospahte nucleotidyltransferase	4e-14	38.4	Thermus phage (AZU97663)	§
	21	Helix-turn-helix domain-containing protein	4e-16	30.5	*Thioclava indica* (WP_081847113)	
	22	-	-	-	-	
	23	-	-	-	-	
	24	-	-	-	-	
	25	-	-	-	-	
SV4	1	-	-	-	-	
	2	Hypothetical protein YSLV6_ORF29	2e-18	28.9	YSLV6 (YP_009177844)	**
	3	Putative ftsk-hera family atpase	4e-81	51.2	QLV (AIF72167)	**
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	Hypothetical protein YSLV6_ORF11	1e-64	44.6	YSLV6 (YP_009177826)	**
	8	-	-	-	-	
	9	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV4	10	Putative cysteine protease	4e-31	42.5	QLV (AIF72172)	**
(contd)	11	Hypothetical protein YSLV6_ORF09	2e-23	40.2	YSLV6 (YP_009177824)	**
	12	SET domain-containing protein	5e-14	28.9	unc. Chlorobi (RMF36440)	
	13	Hypothetical protein	5e-06	37.1	*Trueperella pyogenes* (WP_126919898)	
	14	Tail fiber protein	3e-05	36.5	*Catalinimonas alkaloidigena* (WP_089678304)	
	15	-	-	-	-	
	16	Putative minor capsid protein	2e-76	41.1	QLV (AIF72184)	**
	17	Putative major capsid protein	0	53.1	QLV (AIF72183)	**
	18	Hypothetical protein QLV_16	1e-09	26.8	QLV (AIF72182)	**
	19	-	-	-	-	
	20	Hypothetical protein YSLV6_ORF20	5e-22	35.8	YSLV6 (YP_009177835)	**
	21	-	-	-	-	
	22	-	-	-	-	
	23	-	-	-	-	
	24	-	-	-	-	
	25	Hypothetical protein	8e-33	48	*Criblamydia sequanensis* (WP_041017161)	
SV5	1	Hypothetical protein	2e-05	43.2	*Klebsiella pneumoniae* (WP_107342058)	
	2	Putative cysteine protease	3e-24	39.5	QLV (AIF72172)	**
	3	Hypothetical protein YSLV6_ORF09	2e-24	42.7	YSLV6 (YP_009177824)	**
	4	-	-	-	-	
	5	Class I SAM-dependent methyltransferase	1e-11	33.9	*Nitrosomonas* sp. (WP_107803427)	
	6	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV5	7	-	-	-	-	
(contd)	8	-	-	-	-	
	9	Hypothetical protein YSLV6_ORF11	2e-65	44.8	YSLV6 (YP_009177826)	**
	10	-	-	-	-	
	11	-	-	-	-	
	12	Hypothetical protein YSLV6_ORF29	8e-11	24.3	YSLV6 (YP_009177844)	**
	13	Putative ftsk-hera family atpase	5e-82	53.1	QLV(AIF72167)	**
	14	-	-	-	-	
	15	Uncharacterized protein LOC106599311 isoform X1	2e-11	37.6	*Salmo salar* (XP_014045953)	
	16	-	-	-	-	
	17	Hypothetical protein	2e-04	33.3	*Aliivibrio fischeri* (AKN38939)	
	18	-	-	-	-	
	19	Hypothetical protein	4e-29	27.8	*Phaeodactylum tricornutum* (XP_002180933)	
	20	Hypothetical protein AURANDRAFT_63769	6e-41	23.5	*Aureococcus anophagefferens* (XP_009036825)	
SV6	1	-	-	-	-	
	2	Hypothetical protein	1e-33	44.5	*Ectocarpus siliculosus* (CBJ48608)	
	3	Hypothetical protein PHYSODRAFT_459935, partial	1e-14	40.1	*Phytophthora sojae* (XP_009533852)	
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	-	-	-	-	
	8	GDP-L-fucose synthetase	2e-46	65.2	*Trypanosoma theileri* (ORC88424)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV6	9	Putative GDP-L-fucose synthetase	4e-106	67.8	*Coccomyxa subellipsoidea* (XP_005647335)	
(contd)	10	Tubulin polyglutamylase TTLL5 isoform X2	1e-18	33	*Poecilia formosa* (XP_016525585)	
	11	-	-	-	-	
	12	Yubulin polyglutamylase TTLL4-like	2e-13	29.8	*Sinocyclocheilus anshuiensis* (XP_016342000)	
	13	Dnaj family protein	7e-25	39.5	*Hondaea fermentalgiana* (GBG26159)	
	14	-	-	-	-	
	15	Hypothetical protein PISMIDRAFT_165588	9e-04	44.2	*Pisolithus microcarpus* (KIK27685)	
	16	Twinkle protein	3e-66	54.3	*Thraustotheca clavata* (OQR96724)	
	17	Helicase twinkle	1e-111	47	*Thecamonas trahens* (XP_013758547)	
	18	Hypothetical protein THAPSDRAFT_26082	5e-44	45.5	*Thalassiosira pseudonana* (XP_002286935)	
	19	Hypothetical protein GPECTOR_54g246	2e-53	52.4	*Gonium pectorale* (KXZ45504)	
SV7	1	Hypothetical protein	1e-87	45.4	unc. Mediterranean phage (BAR35703)	§
	2	Ribonucleoside-diphosphate reductase, adenosylcobalamin-dependent	0	74.2	Candidatus *Pelagibacter* sp. (OUU62749)	
	3	Hypothetical protein CBC55_00045	2e-42	53.2	unc. Gammaproteobacteria (OUV23883)	
	4	Hypothetical protein	2e-06	69.4	unc. Mediterranean phage (BAQ85098)	§
	5	-	-	-	-	
	6	-	-	-	-	
	7	Hypothetical protein MTPG_00033	2e-59	75.2	Methylophilales phage (AFB70784)	§
	8	-	-	-	-	
	9	Hypothetical protein CBC65_010385	2e-55	40.2	unc. Rhodothermaceae (RPF78496)	
	10	Hypothetical protein CBD24_02970	6e-06	44.8	unc. Euryarchaeota (OUW13684)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV7	11	Hypothetical protein phage1322_26	4e-18	67.9	Puniceispirillum phage (YP_008320288)	§
(contd)	12	Hypothetical protein	1e-15	37.9	*Leptolyngbya* sp. (WP_068385498)	
	13	Hypothetical protein	1e-31	53.2	unc. Mediterranean phage (BAQ88147)	§
	14	Hypothetical protein	3e-15	40.5	unc. Mediterranean phage (BAR36310)	§
	15	Hypothetical protein	3e-61	47.9	unc. Mediterranean phage (ANS04980)	§
	16	Hypothetical protein	9e-36	45.5	unc. marine virus (AKH46294)	*
	17	Hypothetical protein	4e-70	69.9	unc. Mediterranean phage (BAR35715)	§
	18	Hypothetical protein	7e-10	54.1	unc. Mediterranean phage (BAQ89706)	§
	19	Hypothetical protein	1e-58	42.9	*Pseudomonas entomophila* (WP_044488363)	
	20	Hypothetical protein CBC71_06035	3e-28	41	unc. Rhodobacteraceae (OUV41230)	
	21	Hypothetical protein	2e-07	35.4	unc. Mediterranean phage (BAR15364)	§
	22	Filamentous hemagglutinin-like protein	2e-07	30.9	unc. Mediterranean phage (BAR18064)	§
	23	Hypothetical protein	2e-27	52.5	unc. Mediterranean phage (BAR18205)	§
	24	Virion structural protein	3e-46	68.4	unc. Mediterranean phage (ANS04576)	§
	25	-	-	-	-	
	26	Hypothetical protein	1e-61	57.9	unc. Mediterranean phage (BAR35721)	§
	27	Hypothetical protein	2e-12	58.5	unc. Mediterranean phage (BAR35721)	§
	28	Hypothetical protein	2e-14	58.9	unc. Mediterranean phage (BAR35722)	§
	29	Hypothetical protein	1e-15	47.5	unc. Mediterranean phage (BAR35723)	§
	30	-	-	-	-	
	31	Major head protein	0	70.3	unc. Mediterranean phage (BAR35726)	§
	32	Hypothetical protein	8e-76	52.3	unc. Mediterranean phage (BAR35727)	§

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV8	1	-	-	-	-	
	2	Hypothetical protein EMIHUDRAFT_438276	2e-42	36.7	*Emiliania huxleyi* (XP_005761050)	
	3	Hypothetical protein B4U79_16983	1e-08	42.4	*Dinothrombium tinctorium* (RWR99703)	
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	Hypothetical protein trisph2_011690	5e-15	27.7	T*richoplax* sp. (RDD36227)	
	8	-	-	-	-	
	9	PGV PGCG_00042-like protein	6e-10	33.3	*Phaeocystis globosa* virophage (YP_008059899)	**
	10	Hypothetical protein	3e-04	42	*Aquimarina macrocephali* (WP_024770661)	
	11	Hypothetical protein	3e-06	46.7	*Sinorhizobium meliloti* (WP_088194312)	
	12	-	-	-	-	
	13	Hypothetical protein B4U80_13961	1e-28	31.5	*Leptotrombidium deliense* (RWS23085)	
	14	DNA primase	8e-21	28	Harvfovirus sp. (AYV81568)	*
	15	-	-	-	-	
	16	Putative primase-helicase	1e-20	29.4	YSLV6 (YP_009177818)	**
	17	-	-	-	-	
	18	D5 family helicase-primase	1e-25	30	*Bodo saltans* virus (ATZ80378)	*
	19	-	-	-	-	
SV9	1	Hypothetical protein LOC109474722	2e-09	39.5	*Branchiostoma belcheri* (XP_019630644)	
	2	-	-	-	-	
	3	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV9	4	-	-	-	-	
(contd)	5	-	-	-	-	
	6	-	-	-	-	
	7	Putative ftsk-hera family atpase	3e-81	52.7	QLV (AIF72167)	**
	8	Hypothetical protein YSLV6_ORF29	6e-10	23.9	YSLV6 (YP_009177844)	**
	9	-	-	-	-	
	10	-	-	-	-	
	11	Hypothetical protein YSLV6_ORF11	2e-65	44.8	YSLV6 (YP_009177826)	**
	12	-	-	-	-	
	13	-	-	-	-	
	14	-	-	-	-	
	15	Class I SAM-dependent methyltransferase	1e-11	33.9	*Nitrosomonas* sp. (WP_107803427)	
	16	-	-	-	-	
	17	Hypothetical protein YSLV6_ORF09	2e-24	42.7	YSLV6 (YP_009177824)	**
	18	Putative cysteine protease	3e-24	39.5	QLV (AIF72172)	**
	19	Hypothetical protein	1e-06	33.3	*Skermanella stibiiresistens* (WP_051513603)	
	20	Hypothetical protein	1e-05	27.2	*Campylobacter concisus* (WP_107956208)	
	21	Tail fiber protein	1e-04	37.3	*Odoribacter* sp. (WP_118774300)	
	22	Putative minor capsid protein	3e-52	39.3	QLV (AIF72184)	**
SV10	1	-	-	-	-	
	2	-	-	-	-	
	3	Class I SAM-dependent methyltransferase	1e-11	33.9	*Nitrosomonas* sp. (WP_107803427)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV10	4	-	-	-	-	
(contd)	5	Hypothetical protein YSLV6_ORF09	2e-24	42.7	YSLV6 (YP_009177824)	**
	6	Putative cysteine protease	3e-24	39.5	QLV (AIF72172)	**
	7	Tail fiber protein	5e-05	32.1	unc. Muribaculaceae (WP_123407951)	
	8	Hypothetical protein	1e-05	27.2	*Campylobacter concisus* (WP_107956208)	
	9	Tail fiber protein	1e-04	37.3	*Odoribacter* sp. (WP_118774300)	
	10	Putative minor capsid protein	3e-80	41.6	QLV (AIF72184)	**
	11	Putative major capsid protein	0	53.4	QLV (AIF72183)	**
	12	-	-	-	-	
	13	Hypothetical protein QLV_16	5e-08	52.7	QLV (AIF72182)	**
	14	-	-	-	-	
	15	Hypothetical protein QLV_13	6e-21	33.2	QLV (AIF72179)	**
	16	-	-	-	-	
	17	-	-	-	-	
	18	Collagen-like protein	9e-53	58.2	*Thalassotalea crassostreae* (WP_068547606)	
	19	Collagen-like protein	6e-74	56.8	*Thalassotalea crassostreae* (WP_068547606)	
SV11	1	Putative protein-primed B-family DNA polymerase	0	95.5	Maverick-related virus (YP_004300281)	**
	2	Hypothetical protein	4e-74	97.3	Maverick-related virus (YP_004300282)	**
	3	Hypothetical protein	1e-55	98.9	Maverick-related virus (YP_004300283)	**
	4	Hypothetical protein	5e-91	98.5	Maverick-related virus (YP_004300284)	**
	5	Hypothetical protein	1e-76	96	Maverick-related virus (YP_004300286)	**
	6	Hypothetical protein	2e-131	100	Maverick-related virus (YP_004300287)	**

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV11	7	Hypothetical protein	2e-60	100	Maverick-related virus (YP_004300288)	**
(contd)	8	-	-	-	-	
	9	Hypothetical protein	9e-30	98.1	Maverick-related virus (YP_004300289)	**
	10	Hypothetical protein	5e-155	100	Maverick-related virus (YP_004300290)	**
	11	Hypothetical protein	0	99.4	Maverick-related virus (YP_004300291)	**
	12	Hypothetical protein	0	99.3	Maverick-related virus (YP_004300292)	**
	13	Putative ftsk-hera family atpase	0	100	Maverick-related virus (YP_004300293)	**
	14	Putative cysteine protease	5e-131	95.2	Maverick-related virus (YP_004300294)	**
	15	mavirus penton protein	0	99	Cafeteriavirus-dependent mavirus (6G42_A)	**
	16	major capsid protein	0	99	Cafeteriavirus-dependent mavirus (6G45_A)	**
	17	Hypothetical protein	0	98.5	Maverick-related virus (YP_004300297)	**
	18	Hypothetical protein crov528	5e-09	59.2	*Cafeteria roenbergensis* virus (YP_003970161)	*
SV12	1	Hypothetical protein GUITHDRAFT_76774 (partial)	7e-67	38.6	*Guillardia theta* (XP_005826013)	
	2	Hypothetical protein CBB97_12595 (partial)	4e-107	60.6	*Candidatus Endolissoclinum* sp. (OUU23913)	
	3	-	-	-	-	
	4	TATA-box-binding protein 2-like	7e-25	60	*Eurytemora affinis* (XP_023326917)	
	5	-	-	-	-	
	6	-	-	-	-	
	7	-	-	-	-	
	8	-	-	-	-	
	9	-	-	-	-	
	10	Hypothetical protein Ctob_015440, partial	1e-44	49	*Chrysochromulina* sp. (KOO31866)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV12	11	-	-	-	-	
(contd)	12	Sugar-phospahte nucleotidyltransferase	2e-20	37.4	Thermus phage (AZU97663)	§
	13	Helix-turn-helix domain-containing protein	2e-15	34.1	*Paracoccus lutimaris* (WP_114350724)	
	14	-	-	-	-	
SV13	1	Putative ftsk-hera family ATPase	2e-12	39.2	QLV (AIF72167)	**
	2	Hypothetical protein YSLV6_ORF29	8e-18	29.3	YSLV6 (YP_009177844)	**
	3	Predicted protein	9e-05	24.6	*Micromonas commoda* (XP_002507994)	
	4	Adenine specific DNA methyltransferase	6e-131	69.1	*Phaeocystis globosa* virus (YP_008052748)	*
	5	Hypothetical protein RHOBADRAFT_51759	4e-07	34.4	*Rhodotorula graminis* (XP_018272811)	
	6	Putative DNA helicase/primase/polymerase	2e-57	26.1	QLV (AIF72188)	**
	7	-	-	-	-	
	8	Ribosome assembly 4 (RSA4)	7e-52	40.9	*Brachionus plicatilis* (RNA00824)	
	9	-	-	-	-	
	10	-	-	-	-	
	11	-	-	-	-	
SV14	1	-	-	-	-	
	2	-	-	-	-	
	3	-	-	-	-	
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	Hypothetical protein COB29_11450	4e-17	31.5	*Sulfitobacter* sp. (PHR05679)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV14	8	-	-	-	-	
(contd)	9	-	-	-	-	
	10	-	-	-	-	
	11	-	-	-	-	
	12	Putative DNA primase/polymerase	2e-60	27.3	OLV (ADX05784)	**
	13	-	-	-	-	
	14	-	-	-	-	
SV15	1	-	-	-	-	
	2	Replication factor C subunit 1, putative	1e-26	38.4	*Plasmodium yoelii* (XP_724804)	
	3	Hypothetical protein	5e-84	51.3	*Ectocarpus siliculosus* (CBN77287)	
	4	Replication factor C subunit 1	9e-29	33.7	*Lates calcarifer* (XP_018551354)	
	5	-	-	-	-	
	6	-	-	-	-	
	7	DNA primase	3e-38	37.6	Hokovirus HKV1 (ARF10447)	*
	8	Putative primase-helicase	2e-22	23.7	YSLV6 (YP_009177818)	**
	9	-	-	-	-	
	10	Hypothetical protein DRH13_00060	4e-32	42.6	Candidatus Woesebacteria (RLC33072)	
	11	Integrase	4e-11	34.5	*Polaribacter reichenbachii* (WP_068357690)	
SV16	1	-	-	-	-	
	2	DNA topoisomerase II	2e-44	45.2	*Thraustotheca clavata* (OQR97211)	
	3	DNA topoisomerase, type IIA, conserved site	5e-28	66.3	*Nannochloropsis gaditana* (EWM24585)	
	4	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)
SV16	5	DNA topoisomerase II	8e-63	73.1	*Naegleria gruberi* (XP_002682923)
(contd)	6	Hypothetical protein P175DRAFT_0559943	1e-24	27	*Aspergillus ochraceoroseus* (PTU18044)
	7	Curved DNA-binding protein	3e-18	33.2	*Strigomonas culicis* (EPY23564)
	8	-	-	-	-
	9	Uracil phosphoribosyltransferase	2e-06	62.9	*Ancylobacter aquaticus* (WP_126278802)
	10	Uracil phosphoribosyltransferase	4e-22	58.1	*Saprolegnia diclina* (XP_008604280)
	11	Hypothetical protein PC110_g1571	9e-24	49.5	*Phytophthora cactorum* (RAW42277)
	12	-	-	-	-
	13	Hypothetical protein PINS_007390	2e-50	33.5	*Pythium insidiosum* (GAX99537)
	14	-	-	-	-
	15	Hypothetical protein H257_11490	1e-37	72.1	*Aphanomyces astaci* (XP_009836749)
	16	Vacuolar protein sorting-associated protein 25	6e-27	37.2	*Nannochloropsis gaditana* (EWM27184)
	17	Unnamed protein product	2e-27	45.3	*Albugo candida* (CCI50154)
	18	-	-	-	-
SV17	1	Hypothetical protein BCR36DRAFT_411040	3e-09	41.7	*Piromyces finnis* (ORX53901)
	2	Replication factor C subunit 1	2e-14	41.5	*Zea mays* (ONM26978)
	3	Hypothetical protein SDRG_04085	0	63.5	*Saprolegnia diclina* (XP_008607965)
	4	Hypothetical protein BBJ28_00012612	1e-84	60	*Nothophytophthora* sp. (RLN80359)
SV18	1	Hypothetical protein LY90DRAFT_669165	2e-14	44.9	*Neocallimastix californiae* (ORY57320)
	2	-	-	-	-
	3	-	-	-	-
	4	-	-	-	-

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)
SV18	5	Hypothetical protein COB29_11450	8e-19	31.3	*Sulfitobacter* sp. (PHR05679)
(contd)	6	-	-	-	-
	7	Hypothetical protein BDEG_23179	2e-43	33.3	*Batrachochytrium dendrobatidis* (OAJ39321)
	8	-	-	-	-
	9	Putative DNA polymerase	0	38.4	*Exaiptasia pallida* (KXJ24574)
	10	-	-	-	-
SV19	1	Hypothetical protein Ctob_009710	2e-04	24.2	*Chrysochromulina* sp. (KOO34328)
	2	Similar to Ankyrin repeat domain-containing protein 50	9e-18	39.7	*Pyronema omphalodes* (CCX31168)
	3	Hypothetical protein AURANDRAFT_1068, partial	7e-78	34.9	*Aureococcus anophagefferens* (XP_009035358)
	4	Hypothetical protein THAOC_18333	4e-55	35.8	*Thalassiosira oceanica* (EJK61219)
	5	Hypothetical protein	0	63.2	*Monosiga brevicollis* (XP_001744923)
SV20	1	Hypothetical protein EMIHUDRAFT_199717	2e-28	34.4	*Emiliania huxleyi* (XP_005793805)
	2	Hypothetical protein EMIHUDRAFT_100904	3e-23	34.2	*Emiliania huxleyi* (XP_005777626)
	3	Hypothetical protein EMIHUDRAFT_206065	3e-04	38.5	*Emiliania huxleyi* (XP_005778294)
	4	-	-	-	-
	5	-	-	-	-
	6	-	-	-	-
	7	-	-	-	-
	8	-	-	-	-
SV21	1	-	-	-	-
	2	-	-	-	-

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV21	3	Hypothetical protein BCR35DRAFT_87205	8e-13	29.5	*Leucosporidium creatinivorum* (ORY81252)	
(contd)	4	Putative DNA helicase/primase/polymerase	1e-34	39.2	QLV (AIF72188)	**
	5	Hypothetical protein CALVIDRAFT_602699	7e-11	25.7	*Calocera viscosa* (KZO90812)	
	6	Hypothetical protein CVT24_011455	3e-12	25.7	*Panaeolus cyanescens* (PPR02227)	
	7	-	-	-	-	
	8	-	-	-	-	
SV22	1	-	-	-	-	
	2	-	-	-	-	
	3	-	-	-	-	
	4	Hypothetical protein BZG36_01677	2e-05	37.4	*Bifiguratus adelaidae* (OZJ05543)	
	5	MIGE-like protein	2e-11	37.7	*Chrysochromulina ericina* virus (YP_009173512)	*
	6	-	-	-	-	
	7	Hypothetical protein DRH24_16275	1e-11	36.6	unc. Deltaproteobacteria (RLB77419)	
	8	-	-	-	-	
	9	-	-	-	-	
	10	Hypothetical protein ATN89_17420	1e-13	32.1	*Comamonas thiooxydans* (OAD82862)	
	11	-	-	-	-	
SV23	1	-	-	-	-	
	2	Hypothetical protein AURANDRAFT_23143 (partial)	6e-136	66.3	*Aureococcus anophagefferens* (XP_009035289)	
	3	Class I SAM-dependent methyltransferase	7e-33	39.2	*Lechevalieria* (WP_109630579)	
	4	-	-	-	-	
	5	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV23	6	-	-	-	-	
(contd)	7	G1/S-specific cyclin-D2	3e-26	34	*Hondaea fermentalgiana* (GBG29736)	
SV24	1	Putative minor capsid protein	9e-20	53	QLV (AIF72184)	**
	2	Tail fiber protein	2e-05	35	*Odoribacter* sp. (WP_118774300)	
	3	VCBS repeat-containing protein, partial	0	48.7	*Phaeodactylibacter xiamenensis* (WP_044224045)	
	4	Hypothetical protein YSLV6_ORF09	6e-22	36.4	YSLV6 (YP_009177824)	**
	5	Putative cysteine protease	1e-24	33.3	YSLV6 (YP_009177825)	**
SV25	1	-	-	-	-	
	2	Hypothetical protein QLV_03	2e-35	42.5	QLV (AIF72169)	**
	3	Hypothetical protein B7954_05215, partial	3e-07	40.6	*Vibrio cholerae* (ORP61686)	
	4	Hypothetical protein	1e-09	37.2	*Sphaerotilus natans* (WP_051631941)	
	5	-	-	-	-	
	6	Hypothetical protein EOP48_03510	5e-22	27.7	unc. Sphingobacteriales (RYE58453)	
	7	-	-	-	-	
	8	DNA primase	3e-28	26.2	*Phaeocystis globosa* virophage (YP_008059889)	**
SV26	1	-	-	-	-	
	2	-	-	-	-	
	3	Hypothetical protein COB29_11450	3e-18	31.3	*Sulfitobacter* sp. (PHR05679)	
	4	-	-	-	-	
	5	Hypothetical protein BDEG_23179	7e-38	31.8	*Batrachochytrium dendrobatidis* (OAJ39321)	
	6	Uncharacterized protein LOC111058677	2e-97	42.7	*Nilaparvata lugens* (XP_022201924)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV26	7	Uncharacterized protein LOC108743448	1e-17	50	*Agrilus planipennis* (XP_018334518.2)	
(contd)	8	Uncharacterized protein LOC111102107	3e-23	45.4	*Crassostrea virginica* (XP_022290467)	
	9	-	-	-	-	
SV27	1	Bspa family leucine-rich repeat surface protein	2e-20	58.2	*Polaribacter* sp. (WP_052107465)	
	2	Predicted protein	2e-05	24.6	*Micromonas commoda* (XP_002507994)	
	3	Hypothetical protein YSLV6_ORF29	2e-18	28.9	YSLV6 (YP_009177844)	**
	4	Putative ftsk-hera family atpase	4e-81	51.2	QLV (AIF72167)	**
	5	-	-	-	-	
	6	Hypothetical protein BXU06_16055	2e-14	67.7	*Aquaspirillum* sp. (AQR66390)	
	7	Hypothetical protein	2e-09	33.6	DLV1 (ALN97656)	**
	8	-	-	-	-	
SV28	1	-	-	-	-	
	2	Chaperonin groel	0	85.5	unc. virus (AQM32683)	*
	3	Co-chaperonin groes	5e-37	73.3	unc. virus (ASN63467)	*
	4	Hypothetical protein CBC30_00115	2e-60	56.8	*Chloroflexi bacterium* (OUU78042)	
	5	-	-	-	-	
	6	Hypothetical protein CBD27_11755, partial	1e-13	56.8	unc. Rhodospirillaceae (OUW23939)	
SV29	1	-	-	-	-	
	2	-	-	-	-	
	3	-	-	-	-	
	4	Putative major capsid protein	0	53.1	QLV (AIF72183)	**
	5	Putative minor capsid protein	2e-82	43.3	QLV (AIF72184)	**

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV29	6	Tail fiber protein	3e-05	45.5	*Odoribacter* sp. (WP_118774300)	
(contd)	7	-	-	-	-	
SV30	1	Hypothetical protein SPPG_00926	5e-06	27.1	*Spizellomyces punctatus* (XP_016611481)	
	2	Hypothetical protein CBB97_26410	4e-26	41.7	candidatus *Endolissoclinum* sp. (OUU13224)	
	3	Hypothetical protein VSDG_03464	3e-04	52.6	*Valsa sordida* (ROW00115)	
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	-	-	-	-	
	8	-	-	-	-	
SV31	1	Hypothetical protein YSLV6_ORF11	2e-39	45.4	YSLV6 (YP_009177826)	**
	2	-	-	-	-	
	3	-	-	-	-	
	4	-	-	-	-	
	5	Hypothetical protein	6e-17	34.5	*Bathycoccus prasinos* (XP_007515515)	
	6	-	-	-	-	
	7	Hypothetical protein DRO61_04575 (partial)	3e-06	50	candidatus Bathyarchaeota (RLI49854)	
	8	Hypothetical protein YSLV6_ORF09	1e-23	41	YSLV6 (YP_009177824)	**
	9	Putative cysteine protease	3e-21	34.4	YSLV6 (YP_009177825)	**
	10	Hypothetical protein	2e-04	41.9	*Epibacterium mobile* (WP_114962031)	
SV32	1	Hypothetical protein TSUD_62050	1e-23	37.7	*Trifolium subterraneum* (GAU36936)	
	2	3-hydroxyacyl-coa dehydrogenase type-2	4e-09	73.7	*Hondaea fermentalgiana* (GBG34250)	
	3	3-hydroxyacyl-coa dehydrogenase type-2	7e-93	62.2	*Hondaea fermentalgiana* (GBG34250)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV32	4	-	-	-	-	
(contd)	5	Hypothetical protein fisn_2Hh414	1e-100	33.1	*Fistulifera solaris* (GAX26651)	
	6	-	-	-	-	
SV33	1	ATP-binding cassette sub-family a member 3	3e-56	45.9	*Chrysochromulina* sp. (KOO44197)	
	2	ATP-binding cassette sub-family A member 3	2e-10	53.4	*Orchesella cincta* (ODM97667)	
	3	ABC transporter A family	0	32.8	*Klebsormidium nitens* (GAQ79910)	
SV34	1	Probable multidrug resistance-associated protein	3e-77	27.1	*Nilaparvata lugens* (XP_022194261)	
	2	Multidrug resistance-associated protein 1	2e-99	39.5	*Dufourea novaeangliae* (XP_015430843)	
SV35	1	Putative DNA topoisomerase II	7e-49	40.9	*Ustilago maydis* (XP_011389948)	
	2	Hypothetical protein	3e-74	59.6	*Micromonas pusilla* (XP_003057506)	
	3	Hypothetical protein THAOC_28668	2e-37	58.8	*Thalassiosira oceanica* (EJK52100)	
	4	-	-	-	-	
	5	Hypothetical protein BBJ29_000473	1e-99	39.7	*Phytophthora kernoviae* (RLN53306)	
	6	-	-	-	-	
	7	Carnitine O-acetyltransferase isoform X2	2e-07	45.5	*Oreochromis niloticus* (XP_019205792)	
SV36	1	-	-	-	-	
	2	Putative ftsk-hera family ATPase	2e-83	54.6	QLV (AIF72167)	**
	3	Hypothetical protein YSLV6_ORF29	3e-18	28.1	YSLV6 (YP_009177844)	**
	4	-	-	-	-	
	5	-	-	-	-	
SV37	1	-	-	-	-	
	2	Hypothetical protein vbbcos136_00037	1e-10	36.1	*Bacillus* phage (AYP68169)	§

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV37	3	Putative cysteine protease	8e-24	32.1	YSLV6 (YP_009177825)	**
(contd)	4	Hypothetical protein YSLV6_ORF09	6e-23	39.4	YSLV6 (YP_009177824)	**
	5	-	-	-	-	
	6	-	-	-	-	
	7	-	-	-	-	
	8	-	-	-	-	
	9	-	-	-	-	
	10	Hypothetical protein QLV_05	6e-25	49.1	QLV (AIF72171)	**
SV38	1	Tail fiber protein	8e-07	37.5	*Meiothermus* sp. (WP_110526007)	
	2	-	-	-	-	
	3	-	-	-	-	
	4	-	-	-	-	
	5	-	-	-	-	
	6	-	-	-	-	
	7	-	-	-	-	
SV39	1	Putative ftsk-hera family atpase	3e-25	60.3	QLV (AIF72167)	**
	2	Hypothetical protein	3e-08	29.3	DLV1 (ALN97676)	**
	3	Putative DNA helicase/primase/polymerase	2e-32	31.8	QLV (AIF72188)	**
SV40	1	ATP-dependent Clp protease ATP-binding subunit	1e-96	63.6	*Cyanobium* sp. (WP_006910091)	
	2	ATP-dependent Clp protease ATP-binding subunit	2e-72	39.1	*Yaniella halotolerans* (WP_022869976)	
	3	Hypothetical protein fisn_26Hh014	8e-37	40.7	*Fistulifera solaris* (GAX18724)	
	4	-	-	-	-	
SV41	1	Hypothetical protein	3e-09	45	*Helicobacter apodemus* (WP_052087261)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV41	2	Hypothetical protein PHYSODRAFT_318236	1e-26	35.3	*Phytophthora sojae* (XP_009534388)	
(contd)	3	-	-	-	-	
	4	-	-	-	-	
	5	Hypothetical protein	7e-07	32.4	*Butyricicoccus porcorum* (WP_087020681)	
	6	Hypothetical protein YSLV5_ORF11	6e-30	35.8	YSLV 5 (YP_009177794)	**
	7	-	-	-	-	
SV42	1	Hypothetical protein BBP00_00003954	5e-52	58.7	*Phytophthora kernoviae* (RLN63690)	
	2	Putative minor histocompatibility antigen 13 isoform 1 isoform 11	8e-25	31.6	*Acanthamoeba castellanii* (XP_004368259)	
	3	-	-	-	-	
	4	Serine/threonine protein kinase	2e-33	44.4	candidatus *Pelagibacter* sp. (OUX38613)	
	5	Glycine amidinotransferase	7e-34	31.2	*Nonomuraea candida* (WP_043635897)	
	6	-	-	-	-	
SV43	1	-	-	-	-	
	2	-	-	-	-	
	3	Hypothetical protein YSLV6_ORF20	4e-22	35.7	YSLV6 (YP_009177835)	**
	4	-	-	-	-	
	5	Hypothetical protein YSLV6_ORF21	8e-09	30.8	YSLV6 (YP_009177836)	**
	6	-	-	-	-	
SV44	1	Mrp-3	9e-83	50.5	*Pristionchus pacificus* (PDM60902)	
	2	Unnamed protein product	2e-43	64.8	*Vitrella brassicaformis* (CEM11400)	
	3	ATP-binding cassette glutathione S-conjugate transporter YCF1	2e-15	49	*Rhizophagus irregularis* (EXX51701)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV44	4	Hypothetical protein K457DRAFT_22549	6e-71	44.4	*Mortierella elongata* (OAQ26146)	
(contd)	5	ABC transporter transmembrane region-domain-containing protein	2e-11	43.2	*Jimgerdemannia flammicorona* (RUS27181)	
SV45	1	Ribose-phosphate pyrophosphokinase 4	1e-22	46.9	*Micractinium conductrix* (PSC75590)	
	2	Ribose-phosphate pyrophosphokinase 4 isoform X2	1e-59	45.1	*Lupinus angustifolius* (XP_019435207)	
	3	Trna (adenine-N(6)-)-methyltransferase	2e-09	31.1	unc. Thermoplasmata (RLF37268)	
SV46	1	Putative minor capsid protein	3e-46	41.1	QLV (AIF72184)	**
	2	Putative major capsid protein	0	53.6	QLV (AIF72183)	**
	3	-	-	-	-	
SV47	1	-	-	-	-	
	2	Hypothetical protein YSLV6_ORF21	8e-09	30.8	YSLV6 (YP_009177836)	**
	3	-	-	-	-	
	4	Hypothetical protein OLV12	2e-23	33.9	OLV (ADX05773)	**
	5	-	-	-	-	
SV48	1	ATP-dependent Clp protease ATP-binding subunit	1e-142	45.3	*Chloroflexi bacterium* (RLC82397)	
	2	Chaperone protein clpc4	9e-07	74.2	*Dichanthelium oligosanthes* (OEL38616)	
SV49	1	-	-	-	-	
	2	Hypothetical protein YSLV6_ORF09	1e-25	43.4	YSLV6 (YP_009177824)	**
	3	Putative cysteine protease	5e-24	39.5	QLV (AIF72172)	**
	4	-	-	-	-	
SV50	1	-	-	-	-	
	2	Hypothetical protein AURANDRAFT_69595	2e-116	61.2	*Aureococcus anophagefferens* (XP_009032296)	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV50 (contd)	3	Hypothetical protein fcc1311_030652	6e-05	37.2	*Hondaea fermentalgiana* (GBG26843)	
SV51	1	-	-	-	-	
	2	-	-	-	-	
	3	-	-	-	-	
	4	Hypothetical protein YSLV6_ORF11	6e-68	46.1	YSLV6 (YP_009177826)	**
	5	-	-	-	-	
SV52	1	-	-	-	-	
	2	Hypothetical protein QLV_16	2e-09	34.5	QLV (AIF72182)	**
	3	-	-	-	-	
	4	Putative major capsid protein	3e-30	54.5	QLV (AIF72183)	**
SV53	1	Hypothetical protein YSLV5_ORF11	2e-29	36.2	YSLV5 (YP_009177794)	**
	2	-	-	-	-	
	3	-	-	-	-	
	4	Hypothetical protein PC110_g7963	8e-18	32.9	*Phytophthora cactorum* (RAW35768)	
SV54	1	Hypothetical protein YSLV5_ORF11	3e-28	35.3	YSLV5 (YP_009177794)	**
	2	-	-	-	-	
	3	-	-	-	-	
	4	Hypothetical protein L917_03129	7e-34	37.6	*Phytophthora parasitica* (ETM00120)	
SV55	1	Hypothetical protein YSLV5_ORF11	2e-29	36.2	YSLV5 (YP_009177794)	**
	2	-	-	-	-	
	3	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV55 (contd)	4	Hypothetical protein PC110_g7963	8e-18	32.9	*Phytophthora cactorum* (RAW35768)	
SV56	1	Putative minor capsid protein	2e-22	45.1	QLV (AIF72184)	**
	2	Tail fiber protein	1e-05	39.2	*Odoribacter* sp. (WP_118774300)	
	3	Hypothetical protein AUJ49_07810	1e-07	34.9	unc. Desulfovibrionaceae (OIO01431)	
	4	-	-	-	-	
SV57	1	-	-	-	-	
	2	-	-	-	-	
	3	Helix-turn-helix domain-containing protein	2e-16	28	*Paracoccus lutimaris* (WP_114350724)	
	4	Hypothetical protein CBC02_011290	8e-16	28.8	unc. Flavobacteriaceae (RPG63228)	
SV58	1	Hypothetical protein EMIHUDRAFT_113956	3e-24	35.5	*Emiliania huxleyi* (XP_005781417)	
	2	Hypothetical protein EMIHUDRAFT_96740	6e-25	34.8	*Emiliania huxleyi* (XP_005761898)	
	3	-	-	-	-	
SV59	1	-	-	-	-	
	2	Hypothetical protein OLV12	1e-23	34.7	OLV (ADX05773)	**
	3	-	-	-	-	
	4	Hypothetical protein QLV_16	3e-10	55.4	QLV (AIF72182)	**
SV60	1	-	-	-	-	
	2	Tail fiber protein	4e-08	37.1	*Flavobacterium* sp. (WP_045968641)	
	3	Tail fiber protein	3e-06	43.1	*Odoribacter* sp. (WP_118774300)	
	4	Putative minor capsid protein	1e-19	53	QLV (AIF72184)	**
SV61	1	-	-	-	-	

Table S4. Continuation.

Viral contig	ORF n°	Definition/putative protein function	E-value	% identity	Best BLAST hit (GenBank Accession number)	
SV61	2	-	-	-	-	
(contd)	3	DNA helicase ATP-dependent	9e-28	37.2	*Chlorella sorokiniana* (PRW56986)	
	4	DNA helicase ATP-dependent	8e-24	42.7	*Chlorella sorokiniana* (PRW56986)	
SV62	1	Hypothetical protein	2e-07	45.9	*Epibacterium mobile* (WP_114962031)	
	2	Putative cysteine protease	2e-07	32.7	YSLV6 (YP_009177825)	**
	3	Hypothetical protein YSLV6_ORF09	3e-22	37.9	YSLV6 (YP_009177824)	**
	4	-	-	-	-	
SV63	1	Hypothetical protein QLV_05	1e-27	51.7	QLV (AIF72171)	**
	2	-	-	-	-	
	3	-	-	-	-	
	4	-	-	-	-	
SV64	1	Putative non-transporter ABC protein	1e-15	66	*Cavenderia fasciculata* (XP_004361100)	
	2	Hypothetical protein AMAG_05873	2e-19	79.2	*Allomyces macrogynus* (KNE60490)	
	3	Hypothetical protein PROFUN_13688	1e-39	61.7	*Planoprotostelium fungivorum* (PRP78455)	

Abbreviations: SAG, Single Amplified Genome; ORF, Open Reading Frame; unc., uncultured/unclassified; contd, continued; YSLV, Yellowstone Lake virophage;. OLV, Organic Lake virophage; QLV, Qinghai Lake virophage; DLV, Dishui lake virophage.
Predicted viral genes without hit in the Genbank database are shown by the symbol (-).

Supplementary figures

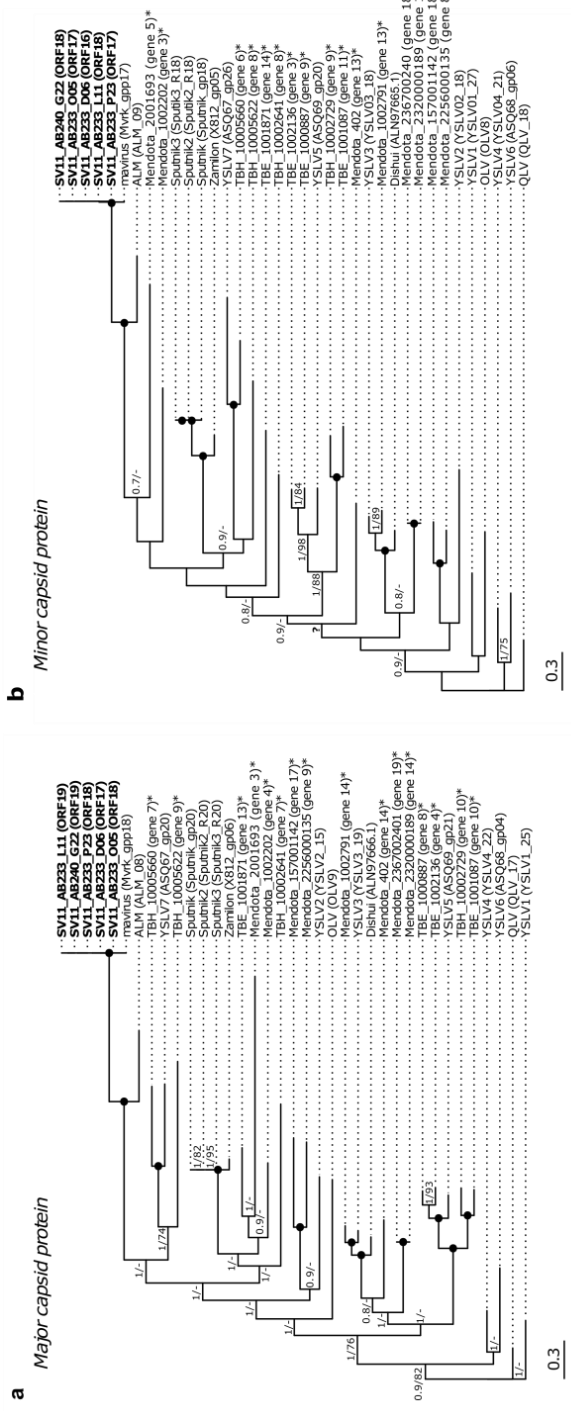


Figure S1. Phylogenetic placement of the putative SAG-associated virophages. The tree topologies were inferred from a maximum-likelihood analysis of genes coding for the **(a)** major capsid protein, **(b)** minor capsid, **(c)** DNA packaging and **(d)** cysteine protease. Bayesian posterior probabilities (BPP) and bootstrap percentages (BS) are provided at each node (BPP/BS) when support values were higher than 0.7 and 70%, respectively. Black dots indicate maximal support for both posterior probabilities (1.0) and maximum-likelihood bootstraps (100%) at the respective nodes. It has to be noted the absence of ATPase gene in TBE_1002136. The five new SAG-associated virophage sequences are highlighted in bold and sequences with different maximum likelihood and bayesian phylogenetic placements are marked by a question mark. For each predicted gene, locus tag or gene name, retrieved in GenBank or in the iPlant Collaborative Discovery Environment (marked by asterisks), respectively, are displayed in parenthesis. Abbreviated name of virophages are detailed in the Material and Methods section. Figure continues in the next page.

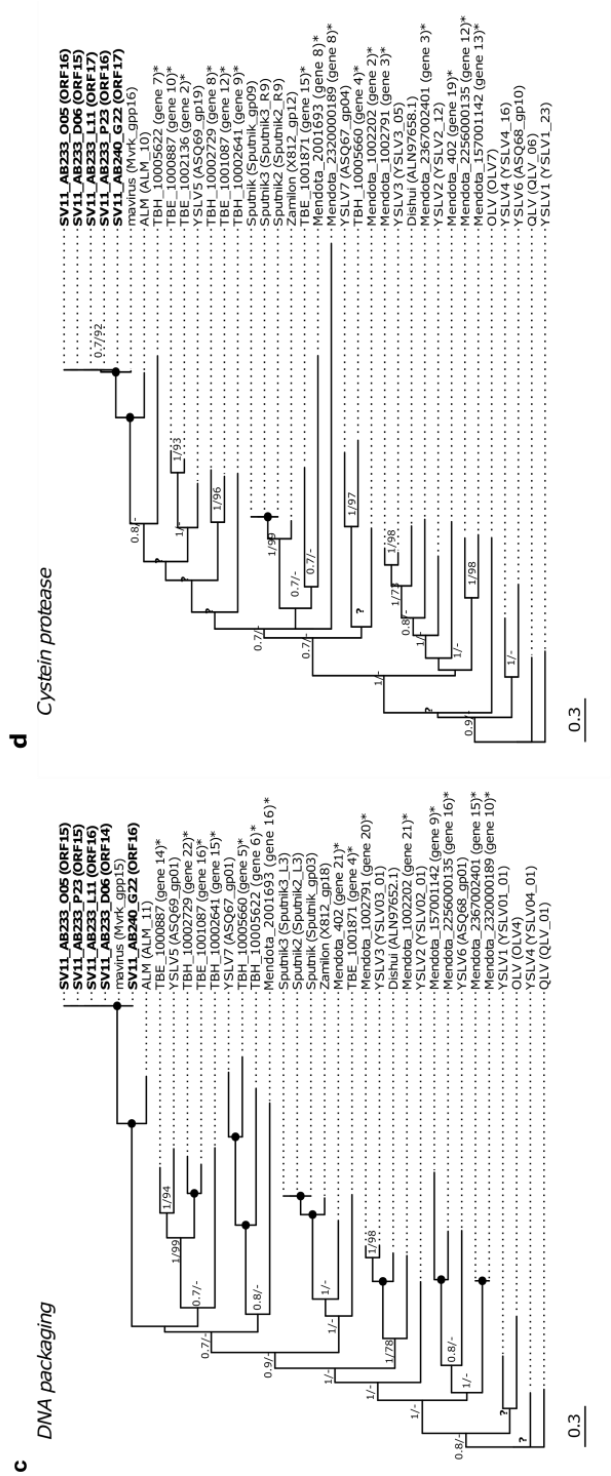


Figure S1. Continuation.

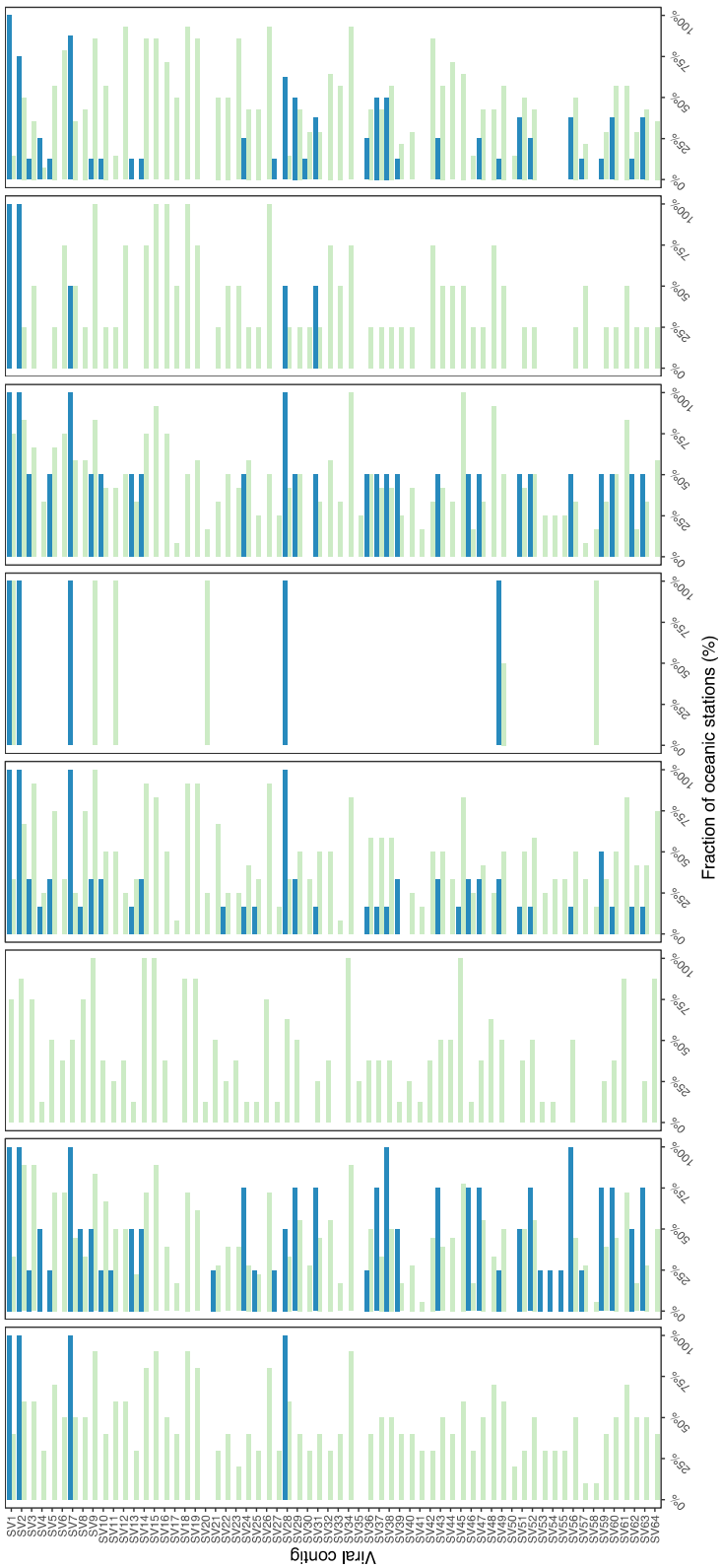


Figure S2. Biogeographical distribution of SAG-associated viruses in the epipelagic zone. Bar plots shown the occurrence of the 64 SAG-associated viral sequences in epipelagic seawater samples (surface and DCM stations combined) among the different crossed oceanic basin. The different basins are as follows: North Pacific Ocean (NP), South Pacific Ocean (SP), North Atlantic Ocean (NA), South Atlantic Ocean (SA), Southern Ocean (SO), Mediterranean Sea (MED), Red Sea (RS) and Indian Ocean (IND). In parenthesis, the total number of stations per basin analyzed among the <0.2µm (left) and 0.2-3µm (right) size fractions.

COMMENTS AND FUTURE PERSPECTIVES

COMMENTS AND FUTURE PERSPECTIVES

Viruses are abundant and dynamic entities in marine ecosystems, and are thought to play a crucial role in controlling the abundances of their hosts, and in the structuring of microbial communities. However, although viral infection on marine microbes (prokaryotes and eukaryotes) has been largely studied through the years (e.g. Suttle *et al.*, 1990; Munn, 2006), it has been mainly done from a bulk perspective, without taking into account that virus–host relationships are rather specific (e.g. Lara *et al*., 2017; Sandaa & Larsen, 2006). This is largely due to the fact that there is not an universal marker to track viruses, and thus knowledge on specific viral–host systems derives mainly from culture studies (e.g. Allers *et al.*, 2013; Derelle *et al.*, 2008; Lara *et al.,* 2015; Vardi *et al.*, 2012). Additionally, most of the work done until now focuses on prokaryotic systems, leaving eukaryotic microbes in the shadows of knowledge (del Campo *et al.*, 2014). Thus, studies addressing virus–host interactions in marine eukaryotes are essential to advance in our understanding of the role of viruses in the ocean.

In this thesis we made an effort to implement and apply techniques that allow the study of interactions between marine picoeukaryotes and their viruses at the single-cell level.

In **Chapter 1** we implemented VirusFISH, which uses fluorescent probes that specifically label both virus and host, to visually follow viral–host interactions. We achieved to monitor the infection dynamics of the virus OtV5 in a non-axenic culture of *Ostreococcus tauri,* unveiling that cell lysis starts much before it is evident from cell counts. Furthermore, VirusFISH enabled the determination of the viral production over time, detecting single free viruses and discriminating them from bacteriophages and other potential artifacts. Finally, VirusFISH also let us to approximate, for the first time, the abundance of viruses in the cellular viral factory prior to the lysis of the cell.

After setting the ground for the use of VirusFISH in a model virus–host system in culture, in **Chapter 2** we applied VirusFISH to study the *Ostreococcus* spp. – virus interactions in nature over a seasonal cycle. Viruses had a variable but notable impact on *Ostreococcus* populations in surface waters along a nearshore to offshore coastal transect, where the percentage of infected cells ranged from 0 to 60%. However no infection could be detected at 50m depth at any time of the year. Although some traditional approaches like the most probable number assay have been used to infer infection dynamics in nature (Cottrell and Suttle, 1995), this is the first time that specific virus–host interactions can be assessed from a visual manner, and the percentage of infected cells can be calculated. We foresee the application of VirusFISH in future studies will expand our knowledge on the impact of viruses in populations of key microbes in the marine environment, which is one of the main unresolved questions in the field of viral ecology.

Finally, in **Chapter 3**, we change the focus from studying defined virus–host systems available in culture to identifying viruses interacting with uncultured marine picoeukaryotes using single-cell genomics. We addressed the viral content of uncultured marine Stramenopiles from diverse lineages, using single amplified genomes (SAG) from the *Tara* Oceans expedition. Even with the low genome recovery, we found that more than half of the cells had viruses associated, suggesting that possibly nearly every cell has a virus. Looking for these viruses in the *Tara* Ocean global metagenomic dataset we were able to establish their biogeography, and showed that some of the viruses were ubiquitous across the global ocean. We also found virophages in two different Stramenopile lineages (chrysophyte-G1 and MAST-3A), highly similar to the mavirus virophage that infects *Cafeteria roenbergensis* (Fischer and Hackl, 2016), which is taxonomically distant from Chrysophyceae and MAST-3. Virophages are thought to act as a defense mechanisms against giant viruses, and our finding suggests that this strategy might be more extended among marine picoeukaryotes than hitherto assumed. Thus, we show that single-cell genomics

is a valuable tool to unveil novel virus–host interactions in picoeukaryotes from a high-throughput perspective, and for formulating hypothesis that can be later tested using other approaches.

The advent of molecular and high-throughput techniques in the field of microbial ecology has provided an unprecedented way to look at processes occurring *in situ* that could not be detected with classical techniques. With the drastic decrease in the sequencing costs experienced in the last years and the improvements in sequencing technologies there is more and more metagenomic information available on viral diversity in marine systems (e.g. Kreuze *et al.*, 2009; López-Pérez *et al.*, 2019). Polymerase chain reaction (PCR) techniques, such as real-time PCR (RT-PCR) (Eggleston and Hewson, 2016), PCR polony method (Baran *et al.*, 2018), digital PCR (Gilg *et al.*, 2016) and droplet digital PCR (Martinez-Hernandez *et al.*, 2019) have been implemented to quantify specific virus groups, providing interesting information on dynamics and biogeography of marine viruses. There is currently even the possibility to sequence genomes of single viruses without the need of culturing (e.g. Lasken, 2012; Labonté *et al.*, 2015; Flores-Uribe *et al.*, 2019). All these technological developments have boosted our knowledge on viral diversity and the role of viruses in the ocean, although they have been mostly applied for the study of bacteriophages.

Our findings on *in situ* marine picoeukaryotic host–virus relationships highlight the need to further expand our knowledge on this compartment of marine microbial communities, which until now is quite unexplored, except for some phototrophic organisms like *Micromonas* (Cottrell and Suttle, 1995; Zingone *et al.*, 1999). Most of the viruses we detected in the SAGs of marine stramenopiles (Chapter 3), which are important members of picoplankton communities, could not be taxonomically classified, likely because very few genomes of viruses infecting marine eukaryotes are available in the databases. Thus, it is important to make the scientific community aware that more effort should be invested on obtaining genomes of eukaryotic viruses. Since most picoeukaryotes are not

amenable to culturing, this could be achieved by increasing the generation of SAGs. The knowledge of more viral genomes will allow us i) to annotate viral sequences detected in metagenomes and assess their geographical distribution, ii) to interpret metatranscriptomic data that are increasingly available in public databases, iii) to build probes to detect those viruses in nature and follow their infection dynamics through VirusFISH. All these approaches will help us obtain a better picture about the ecology of marine viruses and, bringing all these pieces together, we will be able to elucidate the role of eukaryotic viruses in the ocean and build the puzzle of the marine ecology at a better resolution.

REFERENCES

Allers, E., Moraru, C., Duhaime, M.B., Beneze, E., Solonenko, N., Barrero-Canosa, J., et al. (2013) Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environ. Microbiol.* **15**: 2306–2318.

Baran, N., Goldin, S., Maidanik, I., and Lindell, D. (2018) Quantification of diverse virus populations in the environment using the polony method. *Nat. Microbiol.* **3**: 62–72.

del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**: 252–259.

Cottrell, M.T. and Suttle, C.A. (1995) Dynamics of lytic virus infecting the photosynthetic marine picoflagellate *Micromonas pusilla*. *Limnol. Oceanogr.* **40**: 730–739.

Derelle, E., Ferraz, C., Escande, M.-L., Eychenié, S., Cooke, R., Piganeau, G., et al. (2008) Life-cycle and genome of OtV5, a large DNA virus of the pelagic marine unicellular green alga *Ostreococcus tauri*. *PLoS One* **3**: e2250.

Eggleston, E.M. and Hewson, I. (2016) Abundance of two Pelagibacter ubique bacteriophage genotypes along a latitudinal transect in the north and south Atlantic Oceans. *Front. Microbiol.* **7**: 1534.

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

Flores-Uribe, J., Philosof, A., Sharon, I., Fridman, S., Larom, S., and Béjà, O. (2019) A novel uncultured marine cyanophage lineage with lysogenic potential linked to a putative marine *Synechococcus* 'relic' prophage. *Environ. Microbiol. Rep.* **11**: 598–604.

Gilg, I.C., Archer, S.D., Floge, S.A., Fields, D.M., Vermont, A.I., Leavitt, A.H., et al. (2016) Differential gene expression is tied to photo chemical efficiency reduction in virally infected *Emiliania huxleyi*. *Mar. Ecol. Prog. Ser.* **555**: 13–27.

Kreuze, J.F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., and Simon, R. (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. *Virology* **388**: 1–7.

Labonté, J.M., Swan, B.K., Poulos, B., Luo, H., Koren, S., Hallam, S.J., et al. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* **9**: 2386–2399.

Lara, E., Holmfeldt, K., Solonenko, N., Sà, E.L., Ignacio-Espinoza, J.C., Cornejo-Castillo, F.M., et al. (2015) Life-style and genome structure of marine *Pseudoalteromonas* siphovirus B8b isolated from the northwestern Mediterranean Sea. *PLoS One* **10**: 1–26.

Lara, E., Vaqué, D., Sà, E.L., Boras, J.A., Gomes, A., Borrull, E., et al. (2017) Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci. Adv.* **3**: e1602565.

Lasken, R.S. (2012) Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**: 631–640.

López-Pérez, M., Haro-Moreno, J.M., de la Torre, J.R., and Rodriguez-Valera, F. (2019) Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled from metagenomes. *Environ. Microbiol.* **21**: 1980–1988.

Martinez-Hernandez, F., Garcia-Heredia, I., Lluesma Gomez, M., Maestre-Carballa, L., Martínez Martínez, J., and Martinez-Garcia, M. (2019) Droplet Digital PCR for estimating absolute abundances of widespread Pelagibacter viruses. *Front. Microbiol.* **10**: 1226.

Munn, C.B. (2006) Viruses as pathogens of marine organisms—from bacteria to whales. *J. Mar. Biol. Assoc. UK* **86**: 453–467.

Sandaa, R.A. and Larsen, A. (2006) Seasonal variations in virus-host populations in norwegian coastal waters: focusing on the cyanophage community infecting marine *Synechococcus* spp. *Appl. Environ. Microbiol.* **72**: 4610–4618.

Suttle, C.A., Chan, A.M., and Cottrell, M.T. (1990) Infection of phytoplankton by viruses and reduction of primary productivity. *Nature* **347**: 467–469.

Vardi, A., Haramaty, L., Van Mooy, B.A.S., Fredricks, H.F., Kimmance, S.A., Larsen, A., and Bidle, K.D. (2012) Host-virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. *Proc. Natl. Acad. Sci.* **109**: 19327–19332.

Zingone, A., Sarno, D., and Forlani, G. (1999) Seasonal dynamics in the abundance of *Micromonas pusilla* (Prasinophyceae) and its viruses in the Gulf of Naples (Mediterranean Sea). *J. Plankton Res.* **21**: 2143–2159.

GENERAL CONCLUSIONS

GENERAL CONCLUSIONS

Chapter 1

1. The implementation of the VirusFISH technique enables the detection of the *Ostreococcus tauri* - OtV5 virus interaction in culture and to follow its dynamics through all phases of the infection.

2. VirusFISH is a useful tool to measure viral production along the infection occurring in a non-axenic culture, discriminating the eukaryotic viruses from the bacteriophages, or other possible artefacts.

3. The VirusFISH technique allows the estimation of the burst size of the host in non-axenic cultures.

4. Viral probes can be designed to target closely related viruses, allowing the study of the impact of these viruses on lineage-specific populations.

Chapter 2

1. VirusFISH is a powerful method to study the dynamics of the *Ostreococcus* spp. – virus interaction in natural waters.

2. *Ostreococcus* populations were an important component of the picoeukaryotic communities of the Bay of Biscay, particularly in surface waters, where they represented up to 20% of the community.

3. Infection dynamics were variable depending on the station, but the highest proportion of infected cells was detected in July and November-December, when up to 60% of the detected cells were infected.

4. Infection dynamics inferred from VirusFISH were consistent with higher viral transcriptional activity obtained through metatranscriptomics, but VirusFISH had the advantage that allows the calculation of the percentage of infected cells, and thus the impact of virus on a specific population can be estimated.

Chapter 3

1. Single-cell genomics is a valuable approach to study eukaryotic virus–host associations in uncultured hosts. It also allows determining the host range of individual viruses without cultivation.

2. Despite the low genome recovery of the Single Amplified Genomes (SAGs), more than half of the cells had viruses associated, suggesting that there are more viral associations than the ones detected, and possibly a virus in nearly every cell

3. Unlike bacteriophages, which have been reported to be detected in SAGs of different lineages, viruses infecting protists were mostly restricted to one lineage, suggesting they are more specialists than bacteriophages.

4. Only a few of the viral sequences detected in the SAGs could be taxonomically affiliated likely due to the low representation of eukaryotic viruses in genomic databases

5. Fragment recruitment analyses of the viral sequences identified in the SAGs against global ocean metagenomes showed that some viruses were widely distributed, whereas others geographically constrained.

6. The mavirus virophage was detected in two different lineages (chrysophyte-G1 and MAST-3A) that were not hitherto reported to harbor virophages. Thus, virophages seem to be more biologically extended and have a wider host range than previously thought, which hint to the importance of virophages as antiviral defense mechanisms in protist populations.