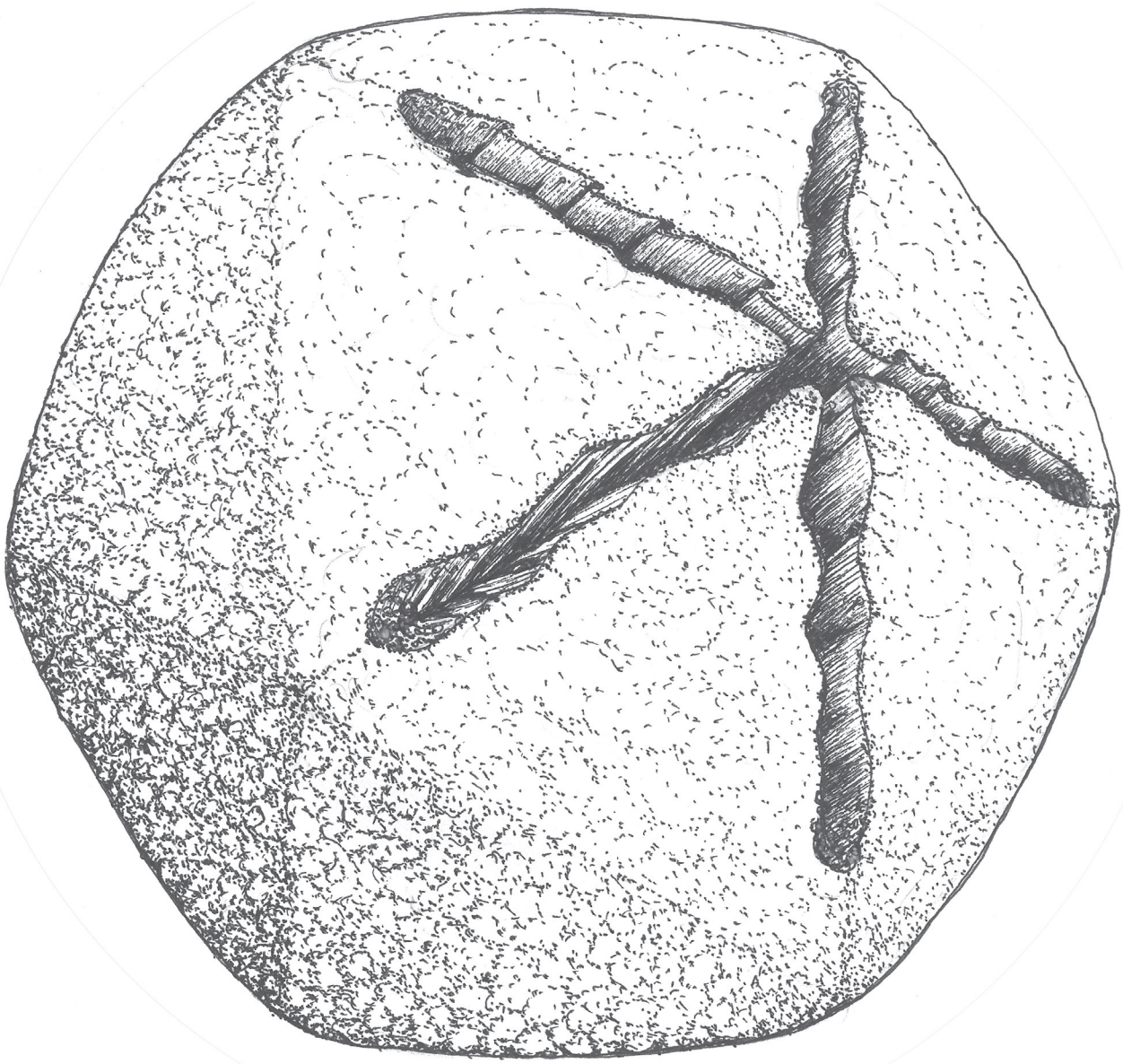
VIRUS DISCOVERY BY METAGENOMICS: THE (IM)POSSIBILITIES

EDITED BY: Bas E. Dutilh, Alejandro Reyes, Richard J. Hall and Katrine L. Whiteson

PUBLISHED IN: Frontiers in Microbiology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88945-308-5
DOI 10.3389/978-2-88945-308-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

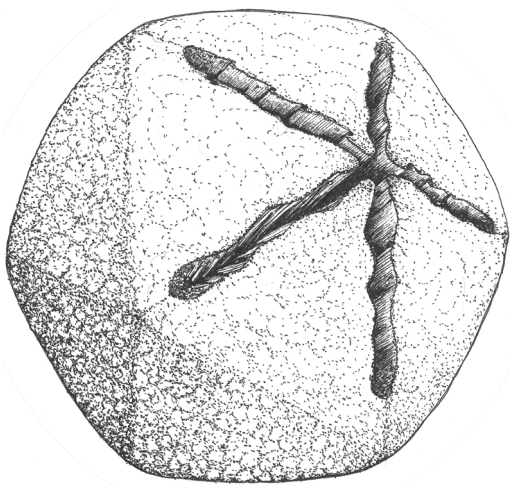
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

VIRUS DISCOVERY BY METAGENOMICS: THE (IM)POSSIBILITIES

Topic Editors:
Bas E. Dutilh, Utrecht University and Radboud University Medical Centre, Netherlands
Alejandro Reyes, Max Planck Tandem Group in Computational Biology, Universidad de los Andes, Colombia; Washington University in Saint Louis, United States
Richard J. Hall, Ministry for Primary Industries, New Zealand
Katrine L. Whiteson, University of California, Irvine, United States



Acanthamoeba polyphaga Mimivirus is a member of the Mimiviridae that packages its huge genome via an aperture in the center of an icosahedral face. The 1,181,549 base pair linear dsDNA genome encodes 979 predicted open reading frames and 39 RNAs. It occurs in freshwater, soil, and host-associated habitats, where it infects its common host, the protozoa Acanthamoeba polyphaga using a lytic lifestyle.

Image: "Life in our phage world" by Forest Rohwer, Merry Youle, Heather Maughan, and Nao Hisakawa (2015phage.org/art.php). Artist: Ben Darby (www.darbyarts.com).

Since the late 1800s, the discovery of new viruses was a gradual process. Viruses were described one by one using a suite of techniques such as (electron) microscopy and viral culture. Investigators were usually interested in a disease state within an organism, and expeditions in viral ecology were rare. The advent of metagenomics using high-throughput sequencing has revolutionized not only the rate of virus discovery, but also the nature of the discoveries. For example, the viral ecology and etiology of many human diseases are being characterized, non-pathogenic viral commensals are ubiquitous, and the description of environmental viromes is making progress. This Frontiers in Virology Research Topic showcases how metagenomic and bioinformatic approaches have been combined to discover, classify and characterize novel viruses.

Citation: Dutilh, B. E., Reyes, A., Hall, R. J., Whiteson, K. L., eds. (2017). Virus Discovery by Metagenomics: The (Im)possibilities. Lausanne: Frontiers Media. doi: 10.3389/978-2-88945-308-5

Table of Contents

Bioinformatic advances driving the new discoveries

frontiers in Microbiology

EDITORIAL
published: 08 September 2017
doi: 10.3389/fmicb.2017.01710

Editorial: Virus Discovery by Metagenomics: The (Im)possibilities

Bas E. Dutilh[1,2]*, *Alejandro Reyes*[3,4,5], *Richard J. Hall*[6] and *Katrine L. Whiteson*[7]

[1] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands, [2] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, Netherlands, [3] Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia, [4] Max Planck Tandem Group in Computational Biology, Universidad de los Andes, Bogotá, Colombia, [5] Center for Genome Sciences and Systems Biology, Washington University in Saint Louis, St Louis, MO, United States, [6] Animal Health Laboratory, Ministry for Primary Industries, Upper Hutt, New Zealand, [7] Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, United States

Keywords: metagenomics, virus discovery, virome, bacteriophages, phages, metagenome, bioinformatics, biological dark matter

Editorial on the Research Topic

Virus Discovery by Metagenomics: The (Im)possibilities

This Frontiers in Virology Research Topic showcases how metagenomic and bioinformatic approaches have been combined to discover, classify and characterize novel viruses. Since the late 1800s (Lecoq, 2001), the discovery of new viruses was a gradual process. Viruses were described one by one using a suite of techniques such as (electron) microscopy and viral culture. Investigators were usually interested in a disease state within an organism, and expeditions in viral ecology were rare. The advent of metagenomics using high-throughput sequencing has revolutionized not only the rate of virus discovery, but also the nature of the discoveries. For example, the viral ecology and etiology of many human diseases are being characterized, non-pathogenic viral commensals are ubiquitous, and the description of environmental viromes is making progress.

This accelerated rate of virus discovery comes both with fantastic possibilities and with significant risks. Metagenomics has already unveiled vast microbial biodiversity in a range of environments, and is increasingly being applied in clinics for difficult-to-diagnose cases. Hall et al. have contributed a thoughtful review on the challenges in using viral metagenomics for diagnostics, including handling of incidental findings, implications for agricultural and horticultural trade, privacy concerns relating to the host's genome, data sharing, cost, quality assurance, and etiology. Presently, the genomic era defines the viral universe by characterizing genotypes, but these genotypes are rarely associated with a phenotype and/or a physical entity. Moreover, the number and diversity of viral sequences in reference databases are dwarfed by the sequences from their cellular hosts. As a result, state of the art taxonomic classification of viruses recognizes only several thousand viral species, a large fraction of which infect humans. This stands in sharp contrast with the diversity of the cellular organisms on which all viruses depend for their replication.

Although fifteen years have passed since the first viral metagenome was sequenced from an ocean sample (Breitbart et al., 2002), the experimental and bioinformatic methods used for viral metagenomics have not reached a consensus. First, this reflects the diversity in applications ranging from virus discovery, to diagnostics, to ecological surveys. Second, this reflects the diversity in the microbial world itself that includes giant viruses (Halary et al., 2016), tiny bacteria (Brown et al., 2015), and everything in between. Thus, even the most basic experimental steps such as filtering viruses from an environmental sample need to be optimized for different applications. Viruses may employ diverse genomic molecular compositions, as illustrated by an RNA sequencing study that uncovered several single stranded and double-stranded RNA viruses in mosquitoes

OPEN ACCESS

Edited by:
Akihide Ryo,
Yokohama City University, Japan

Reviewed by:
Masatsugu Obuchi,
Toyama Institute of Health, Japan

***Correspondence:**
Bas E. Dutilh
bedutilh@gmail.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *04 August 2017*
Accepted: *23 August 2017*
Published: *08 September 2017*

Citation:
Dutilh BE, Reyes A, Hall RJ and Whiteson KL (2017) Editorial: Virus Discovery by Metagenomics: The (Im)possibilities. Front. Microbiol. 8:1710. doi: 10.3389/fmicb.2017.01710

(Chandler et al.). Roux and colleagues studied the uncultivated "Far-T4 phages" that are commonly found in aquatic environments, identifying five clades with largely collinear genome organizations (Roux et al.). Our Research Topic highlights the diversity of the virosphere in reviews on plant viruses where metagenomics has revealed an unexpected diversity in viruses with persistent lifestyles (Roossinck), and on coral reefs where herpes viruses were revealed in various host species (Houldcroft and Breuer).

A third reason for the diversity in viral metagenomics methods is the ongoing development of new tools and approaches, that are both cause and effect of our improved understanding of the virosphere. Bioinformatically, identifying viral sequences in a shotgun metagenomic dataset can be like finding a needle in a haystack (Soueidan et al.). Detection methods based on reference sequences can sensitively identify known viruses in short-read datasets (Pirovano et al.), but may limit the search to identify only known species. A promising possibility is to use short seeds to identify and progressively assemble viral sequences from the dataset, for example allowing the reconstruction of 45 partial or complete *Alpavirinae* genomes (Alves et al.). Alternatively, abundance and nucleotide usage signals can be used to identify *de novo* assembled metagenomic contigs belonging to the same genome, although the specificity of these binning signals varies (Smits et al.). Viral fosmids are a complementary approach allowing the recovery of long contiguous sequences, albeit at the cost of an inherently lower throughput. Chow et al. combined fosmid sequencing with shotgun metagenomics and database searches to chart the viral diversity in a Canadian fjord, elucidating genomic and ecological contexts, and identifying potential host interactions.

It has been known for decades that viruses infect economically important crops and animals including humans. However, a different view of viruses has recently emerged, the virome, a name for the entire community of viruses found in a given biome. For example, the human virome consists of the viruses that normally live on and within a human being, and includes viruses that infect the human itself, but also viruses ingested via food and the viruses infecting human-associated bacteria and archaea. Indeed, viruses are everywhere, and understanding the role of the virome within a complex ecosystem is a challenge at a whole new level. Aziz et al. developed various metrics to represent the presence of sequences in metagenomes and environments, and created a web tool showing the presence of a set of reference genomes in available metagenomes.

In our Research Topic, different approaches were taken. The bacteriophage adherence to mucus (BAM) model, which proposed that bacteriophages could assist the immune system of animal hosts by creating an external layer of defense in mucosal surfaces (Barr et al., 2013), has been extended to wild and farmed eels (Carda-Dieguez et al.). While the BAM model suggests that mainly lytic phages should benefit from this behavior, very similar Ig-like motifs to those originally implicated in mucus attachment were identified in temperate *Pseudomonas aeruginosa* phages (Tariq et al.). Those findings led to a proposed model incorporating BAM into the lifecycle of *P. aeruginosa* in cystic fibrosis patients.

Some viruses in host-associated viromes may chronically linger without causing any symptoms or phenotype, until their emergence is triggered, for example following debilitation of the host immune system in the case of eukaryotic viruses, or other environmental stresses in the case of temperate bacteriophages progressing into the lytic cycle. For example, Santiago-Rodriguez et al. identified a phage in Methicillin Resistant *Staphylococcus aureus* (MRSA) whose expression was inhibited in an *ex vivo* human blood model, suggesting preference for the lysogenic state in blood. The ecological question of virome stability is also relevant outside the context of a host, being linked to organismal diversity and nutrient release (Suttle, 2007). In our Research Topic, marker gene amplification studies targeting *Picornavirales* (Gustavsen et al.) and *Gokushoviruses* (Labonté et al.) in Canadian coastal waters show high viral diversity, both spatiotemporally and across a depth gradient. This variability suggests an important role for the viruses in structuring the bacterial and eukaryotic plankton community, as well as in nutrient cycling and energy transfer.

Santiago-Rodriguez et al. investigated whether the virome could be used as a sensitive marker of alterations in the health status of a host. While the urinary tract was long considered a sterile environment except during rare urinary tract infections (UTIs), it was recently shown that in fact, it contains an associated microbiome even in healthy individuals. Like many human-associated viromes, the urinary tract virome sequences were found to be dominated by bacteriophages. Only 27% of virome contigs were homologous to a known virus (similar to what other studies of human viromes find), and most of the hits matched bacteriophages. No significant changes were detected between healthy individuals and UTI patients (Santiago-Rodriguez et al.). Interestingly, human papillomavirus (HPV) was detected in 95% of the subjects, regardless of disease status. Traditionally, HPV was associated with diseases including cancer, but many HPV genotypes are now widely detected without any apparent association to disease. A similar case is a novel gamma-papillomavirus that was discovered in the virome of a patient with a respiratory tract infection (Canuti et al.). This HPV was present at equivalent titers during the respiratory infection and after the recovery, suggesting it was not the cause of the disease.

Linking unknown viral metagenomic sequencing reads to a function in a complex environment is often impossible, so in vitro systems to study novel phage bacteria interaction could serve as an ideal intermediate. For example, Gambelli et al. characterized the virome in a bioreactor containing a *Methylomirabilis oxyfera* enrichment culture, hoping to identify a phage that infected this important nitrogen cycling bacterium. While the shape and size of the virions could be modeled in high detail by advanced electron microscopy and three-dimensional imaging, and high-throughput metagenomic sequencing identified several very long bacteriophage contigs, it still proved challenging to identify which of the metagenomic sequences represented the phages the *M. oxyfera* phages seen in the images (Gambelli et al.). This impressive effort thus highlights some of the challenges we face on the road ahead toward a full understanding of viruses and their interactions in the natural environment.

CONCLUSIONS AND OUTLOOK

Studying viral sequences means working at the edge of human knowledge. Even microbial genomics experts working on uncultivated microbes use the term "dark matter" when describing the viral sequences they find in metagenomes. While metagenomics expands our ability to detect viruses, a combination of small viral sequence databases and great diversity still means that many viral reads have no homology to known viruses (Mokili et al., 2012). Whether they are host-associated or free-living, we now know that most viruses (like microbes in general) are not pathogenic to humans, plants, or animals. Recent technological advances including decreased DNA sequencing costs and the development of novel methods in metagenome analysis are making the study of viral communities feasible to many laboratories around the world. The Research Topic authors were motivated to identify novel viral agents of disease, illuminate the vast "dark matter" that is viral diversity, discover functional genes carried by bacteriophages, uncover how phages structure microbial communities, and perhaps support renewed interest in phage therapy to target antibiotics resistant bacterial infections. The current Research Topic is an excellent compendium of manuscripts that, far from being comprehensive, we hope will form a foundation and inspiration for many other studies to come in the field of viral discovery, and motivate a new generation of microbial ecologists to include the viruses in their research.

AUTHOR CONTRIBUTIONS

BD, AR, RH, and KW edited the Research Topic and wrote the Editorial.

ACKNOWLEDGMENTS

We wish to thank all authors and reviewers of the manuscripts in this Research Topic "Virus discovery by metagenomics: the (im)possibilities" for their valuable contributions. BD was supported by the Netherlands Organization for Scientific Research NWO Vidi grant 864.14.004.

REFERENCES

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., et al. (2013). Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., et al. (2015). Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* 523, 208–211. doi: 10.1038/nature14486

Halary, S., Temmam, S., Raoult, D., and Desnues, C. (2016). Viral metagenomics: are we missing the giants? *Curr. Opin. Microbiol.* 31, 34–43. doi: 10.1016/j.mib.2016.01.005

Lecoq, H. (2001). Discovery of the first virus, tobacco mosaic virus: 1892 or 1898? *C. R. Acad. Sci. III* 324, 929–933. doi: 10.1016/S0764-4469(01)01368-3

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

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

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

frontiers in Microbiology

HYPOTHESIS AND THEORY
published: 25 March 2015
doi: 10.3389/fmicb.2015.00224

Beyond research: a primer for considerations on using viral metagenomics in the field and clinic

Richard J. Hall[1], Jenny L. Draper[2], Fiona G. G. Nielsen[3] and Bas E. Dutilh[4,5,6]**

[1] Institute of Environmental Science and Research, National Centre for Biosecurity and Infectious Disease, Upper Hutt, New Zealand, [2] Ministry for Primary Industries Animal Health Laboratory, National Centre for Biosecurity and Infectious Disease, Upper Hutt, New Zealand, [3] DNAdigest, Future Business Centre, Cambridge, UK, [4] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands, [5] Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Centre, Nijmegen, Netherlands, [6] Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:
Katrine L. Whiteson,
University of California, Irvine, USA

Reviewed by:
Richard K. Plemper,
Georgia State University, USA
David T. Pride,
University of California, San Diego, USA

***Correspondence:**
Richard J. Hall,
Institute of Environmental Science and Research, National Centre for Biosecurity and Infectious Disease, 66 Ward Street, Upper Hutt 5018, New Zealand
richard.hall@esr.cri.nz;
Bas E. Dutilh,
Theoretical Biology and Bioinformatics, Utrecht University, Padualaan 8, 3584 CH, Utrecht, Netherlands
bedutilh@gmail.com

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: 10 December 2014
Accepted: 06 March 2015
Published: 25 March 2015

Citation:
Hall RJ, Draper JL, Nielsen FGG and Dutilh BE (2015) Beyond research: a primer for considerations on using viral metagenomics in the field and clinic.
Front. Microbiol. 6:224.
doi: 10.3389/fmicb.2015.00224

Powered by recent advances in next-generation sequencing technologies, metagenomics has already unveiled vast microbial biodiversity in a range of environments, and is increasingly being applied in clinics for difficult-to-diagnose cases. It can be tempting to suggest that metagenomics could be used as a "universal test" for all pathogens without the need to conduct lengthy serial testing using specific assays. While this is an exciting prospect, there are issues that need to be addressed before metagenomic methods can be applied with rigor as a diagnostic tool, including the potential for incidental findings, unforeseen consequences for trade and regulatory authorities, privacy and cultural issues, data sharing, and appropriate reporting of results to end-users. These issues will require consideration and discussion across a range of disciplines, with inclusion of scientists, ethicists, clinicians, diagnosticians, health practitioners, and ultimately the public. Here, we provide a primer for consideration on some of these issues.

Keywords: viral metagenomics, ethics, medical, metagenomics, data interpretation, incidental findings, diagnostic tools, biomarkers

Introduction

Efforts to discover and describe new viral species have lagged behind other organisms such as bacteria, fungi, or eukaryotes (Rosario and Breitbart, 2011; Delwart, 2013; Hurwitz and Sullivan, 2013). In part, this is due to the lack of a universal conserved genetic element shared between viral genomes which could be exploited for the purposes of viral genome discovery. For other microbes, there are conserved elements which can be targeted to detect new taxa, such as the 16S ribosomal RNA (rRNA) gene for bacteria, or the internal transcribed spacer (ITS) region for fungi. Next-generation sequencing in shotgun metagenomics has greatly increased the capacity for, and discovery of, new viruses (Mokili et al., 2012). This topic has been the subject of intense scientific effort and review since the first application of this technology in 2002 to describe uncultured viruses in the marine environment (Breitbart et al., 2002). Given the common application of metagenomics for virus discovery in basic and applied research, it is not surprising that there is a growing interest for use in a diagnostic capacity. Potential diagnostic applications of viral metagenomics span many areas, from horticulture (Kehoe et al., 2014) to veterinary medicine (Blomstrom, 2011; Belak et al., 2013) through to human health (Miller et al., 2013). Similarly, metagenomics can be

used as a tool in forensics (Karlsson et al., 2013) and for monitoring environmental samples such as water quality (Ng et al., 2012).

A Universal Test: Incidental Findings

In the example of human health, diagnostic tests traditionally proceed as follows. First, a clinician examines a patient and makes a provisional diagnosis, which may then require further laboratory testing. If an infectious disease is suspected, the clinician is likely to request that pathogen A, B, or C be tested for, and a diagnostic laboratory will use culture, antigen, or antibody detection (immunoassay), or molecular methods. This process is similar for the diagnosis of infectious disease in animals or plants. Syndromic diseases, such as influenza-like respiratory illness in humans, vesicular disease in livestock, or "virus-like symptoms" in plants may require a more complex diagnostic workup, which could take place as a multiplex PCR or parallel/serial specific testing. In general terms, most PCR, culture, or immunoassay-based diagnostic methods only provide evidence for the presence or absence of specific pathogens. This is often all that is required to enable appropriate therapy. Additional techniques may be needed if other attributes of a pathogen are to be determined, i.e., antiviral or antimicrobial resistance, or presence of virulence genes/toxins. In contrast, metagenomics has the potential to detect all known organisms within a sample in a single experiment, and is particularly useful for viruses given the lack of a universally conserved genetic element. Deep metagenomic sequencing has the potential to not only identify viruses with a high level of taxonomic resolution, but may also reveal other attributes that are clinically relevant, such as resistance to antiviral medications (Quinones-Mateu et al., 2014). Because of these advantages, metagenomics is being increasingly applied in diagnostics, and is already being offered as a commercial diagnostic service[1,2].

Thus far, the potential ethical implications of metagenomics to detect all organisms in a sample have received less attention, with a notable absence of discussion in relation to the detection and diagnosis of pathogens. For example, the technique has an inherent capability for incidental detection of pathogens that may not be relevant to the condition for which the patient is seeking treatment. It is therefore prudent to consider the possible impacts of an incidental diagnosis before proceeding with such a test. Consider the hypothetical example of a stool sample sent to a medical laboratory for norovirus testing. If the patient is HIV-positive and metagenomics is applied, then it is likely that HIV sequences will be detected in the metagenomic data (especially relevant for the RNA metagenome, also known as the metatranscriptome), given that up to 60% of HIV patients have detectable levels of HIV in their stool (van der Hoek et al., 1995). Both patient and physician need to be aware of this potential. In cases where such incidental findings are unwanted, a technical solution is possible: using bioinformatics, a panel of high-consequence viruses (or pathogens of concern) could be rapidly filtered from the data before reporting took place.

The potential for incidental findings may seem alarming, but is not unprecedented. Technologies such as magnetic resonance imaging (MRI) or computer-tomography (CT) scanning can detect cancer as an incidental finding, but patients will usually be made aware of this possibility before consenting to a test. The physician will have a clear understanding of how to manage the finding of malignancy, should it be observed. Moreover, we observe a parallel with individualized complete human genome or exome sequencing, where an average human being carries approximately 250–300 loss-of-function variants in annotated genes, and 50–100 variants previously implicated in inherited disorders (Genomes Project et al., 2010). We contend that it is the responsibility of any researcher or diagnostician who uses metagenomics in a clinical setting to ensure that both the physician and patient are aware of the potential for detection of any pathogen, even those unrelated to their disease, and that informed consent information is clear about the processes in place for supporting the patient in case of incidental findings. There are, however, to our knowledge, not yet any established guidelines for how such informed consent should be handled in the case of clinical diagnosis using metagenomic assays, which makes the requirement of informed consent very difficult to implement in practice for individual testing labs.

For clinical exome or whole-genome sequencing of patients, recommendations on incidental findings have been developed by the American College of Medical Genetics and Genomics (ACMG), including: a "minimum list" of serious disease variants which are recommended to be reported from the genetic lab to the ordering clinician; a recommendation on the type of (likely) pathogenic incidental findings to report; guidelines for pre- and post-test counseling of the patient; and recommendations for a patient to opt-out of receiving incidental findings (Green et al., 2013). Guidelines for consent, processing and reporting of incidental findings for clinical metagenomic sequencing addressing the above issues are urgently required to ensure that adequate ethical considerations are met when handling and analyzing human metagenomic data.

The Host Genome: Culture, Ethnicity, and the Law

A metagenomic dataset is likely to contain significant amounts of host genome sequence, depending on the sample type and sequencing protocol. Routine clinical samples in human health such as feces, swabs, tissue biopsy, sputum, and urine, are all likely to contain human DNA and RNA. In some countries, there are strict legal controls in place that govern the use of human tissue. For example, in New Zealand, consent of the patient for the use of their tissue for any future unspecified research purposes is an absolute statutory requirement (Ministry of Health, 2007), and the tissue may only be used for the express intention for which it was collected (New Zealand Government, 2008). In particular, the indigenous people of New Zealand, Māori, have a unique cultural perspective on the ownership

[1]http://www.pathogenica.com/

[2]http://www.aperiomics.com/

of human genetic information, which is not held to be the property of the individual, and decisions about investigating DNA should be made collectively (Baird et al., 1995; Hudson, 2009; New Zealand Health Research Council, 2010). Thus, it is possible that the incidental sequencing of human DNA could lead to legal, ethical, and/or cultural obligations when conducting a metagenomic analysis. Of course, the controls around sequencing of human DNA will vary greatly between jurisdictions, but the implications of generating incidental information should be accounted for. Once again, technical solutions such as bioinformatic filtering of the data for human DNA may provide a solution, but there would need to be some confidence around the quality of this process for deleting human genomic sequences. Without the removal of human sequence, human metagenomic datasets are likely to be subject to legislation for health-related personal data, which implies requirements for subject anonymization and restricted data access (Mascalzoni et al., 2014). To ensure the highest level of sharing and deposition of metagenomic data in research data repositories, we urge repositories and ethical review committees to develop guidelines on which methods are considered adequate for the removal of potentially personally identifiable human sequences from metagenomic data sets, to allow for deposition and sharing of metagenomic data without restrictions. Comprehensive guidelines should take into consideration both the benefit of data usage to inform research and any potential harm for patients as a result of non-compliance to guidelines. Consequences for a breach of ethical approval, and procedures to recall shared data in the event of a breach should also be considered. The ethical considerations for developing governance and guidelines have been discussed in depth by the Nuffield Council on Bioethics in their report concerning the collection and usage of biomedical data for research and health care (Nuffield Council on Bioethics, 2014).

Trade Implications

The detection of pathogens of agricultural or horticultural significance can have dire consequences for productivity and trade. Most nations have government veterinary or plant health laboratories that provide testing and surveillance for significant pathogens. These laboratories specialize in screening animals and goods for trade purposes, preventing the spread of pathogens, and playing a crucial role in demonstrating that a country or region is free of specific diseases. In the case of trade in animals and animal products, the World Organization for Animal Health (OIE) maintains a list of the veterinary diseases of greatest concern worldwide, tracks, and reports on their global occurrence, and sets diagnostic testing standards for these diseases to aid international trade. New outbreaks of OIE-listed diseases in countries previously considered free of the disease can have severe trade implications, resulting in either blockage of relevant trade permits, or a massive increase in the testing and documentation required to certify products for export. The cost of such outbreaks can greatly affect trade-based economies, and proving freedom from such a disease once it is detected (or even suspected) can be a major economic burden.

Due to the sensitivity and untargeted nature of metagenomics, its application to animal or plant samples presents a challenge for regulatory authorities (MacDiarmid et al., 2013). Methods for assigning metagenomic sequencing reads to species may not be perfect, and incomplete reference databases as well as evolutionary conservation between species can easily lead to an incorrect diagnosis. Additionally, genetic databases are still skewed towards heavily researched organisms – particularly pathogens – and this bias increases the chance of relatively benign viruses being classified as "similar to" viruses of significant agricultural or clinical concern. A hypothetical example of a fish metagenome can be used to illustrate this point. Such a metagenome could conceivably contain sequences from a relatively benign virus in the family *Orthomyxoviridae*, which has some genetic similarity to the trade-sensitive *Isavirus* that causes Infectious Salmon Anemia (ISA) in salmon. If such a sequence, found in a healthy fish, is reported as "similar to ISA virus," this finding could have significant implications for trade in salmon, given that even the suggestion that an OIE-listed virus may be present can be enough to impose significant trade restrictions on an exporting country.

Responsibility must be taken by researchers and diagnosticians performing metagenomic analyses to ensure the veracity of their results and to engage with regulatory authorities early on in the process once an initial discovery has been made, to avoid unintended economic harm from falsely declaring the presence of an organism with trade implications.

Use of the Data

It is important that metagenomic datasets are freely and openly shared upon the publication of scientific results, for example by deposition in a database such as EBI Metagenomics (Hunter et al., 2014), MG-RAST (Meyer et al., 2008), or Genbank (Benson et al., 2014). This allows other researchers to investigate the data, to substantiate or refute claims made by the original authors, and enables alternative research projects to be developed for which the samples were not originally obtained. Sharing metagenomic datasets also allows significant scientific discoveries to be made (Dutilh et al., 2014), and restricting access to datasets can seriously delay scientific progress. For example, the Fourth Paradigm of data-driven scientific discovery describes how the advancement of a scientific field depends on how well researchers collaborate with one another (Hey et al., 2009).

However, there are several concerns when it comes to publishing diagnostic metagenomic datasets. For example, as we discuss here, misinterpretation can have serious consequences. Moreover, there is a possibility that pathogens could be missed by the original submitters, or even discovered at a later date.

Additionally, in human diagnostic settings sharing of metagenomic data becomes ethically complex due to issues of patient confidentiality and privacy infringements. For example, if human samples were obtained for a specific purpose like virus discovery, then ethical permissions may only be granted for that purpose alone. If a researcher were to investigate any other aspect of the

metagenomic data – such as looking for a correlation between a human genetic mutation and infection status or outcome – this would be in breach of the original ethical approval and/or statutory obligations.

Expectations of the End-User

Ultimately, a decision must be made on how to act upon the result of a diagnostic test. Such decisions are made every day by clinicians, veterinarians, ecologists, epidemiologists, scientists, farmers, and private citizens. Commercially available tests for pathogen detection often condense a reported result into a binary "presence or absence" call, with much of the technical detail being deliberately and carefully hidden, to make interpretation easy. Metagenomic analyses currently provide end-users with a "shotgun" picture of the microbiome that includes a list of the organisms that are theoretically present based on sequence similarity. Interpretation relies on an expert examining the data and making assessments based on their experience, particularly in regard to the reliability of the methodology, nuances between viral or bacterial strains, genetic similarities between viruses, bacteria, and eukaryotes, and the potential for contamination of genome databases (Gonzalez et al., 2014; Merchant et al., 2014) and nucleic acid extraction kits (Salter et al., 2014). Therefore, the reports currently generated for metagenomic datasets are not yet conducive to widespread use, as witnessed by the recent mass public reporting of the alleged detection of anthrax and the bubonic plague on the New York City subway system, based on a metagenomic analysis which contained potential but non-definitive hits to the these pathogens (Afshinnekoo et al., 2015; Mason, 2015; Yong, 2015).

At least in the US, no tests for human genetic testing are allowed to enter the market as medical devices without strict analytical and clinical validation to ensure consistent and robust results. One recent example is the case of the direct-to-consumer (DTC) human genotyping service provided by the company 23andMe. In November 2013, the 23andMe Personal Genome Service product was banned from marketing and providing "medical reports" of "health risks" and "drug response" when the company failed to deliver adequate evidence to the FDA for validation of specificity and sensitivity (Woods, 2013). Similarly, we see a big challenge for the metagenomics community to develop robust analysis methods before metagenomics can be approved for widespread clinical use. We note that, as of February 2015, the FDA has eased access to DTC DNA screening for several inherited diseases.

Another recent case illustrates the dangers of misinterpretation of metagenomic analysis (based on unpublished data). In this case, an automated metagenomic pipeline, MG-RAST (Meyer et al., 2008) was used to analyze a metagenomic dataset from environmental sample source. Prior to the data analysis, a worker involved in sampling was suffering from an undiagnosed illness. During the preliminary analysis a sequence hit to a Risk Group 3 pathogen was observed in the results from MG-RAST, and the group involved in the sampling suspected that this could be the cause of the worker's illness. When the affected worker requested testing and prophylactic treatment for this notifiable zoonotic disease at a clinic, government organizations in human and animal health became involved. However, upon detailed review of the data, it was revealed that the "sequence hit" in MG-RAST actually matched a known archaeal contaminant in the draft genome assembly of the pathogen, rather than the pathogen itself. Although the contaminant sequences were clearly unrelated to the pathogen and had already been retracted from Genbank, they were still present in the automated annotation pipeline, which relied on an outdated database. We anticipate that such situations are likely to become increasingly common as metagenomics becomes mainstream.

It is our experience that when collaborators are presented with the taxonomic report from a metagenomic study for the first time, they may be overwhelmed. After some time spent studying the data, some even become skeptical. This is due to three factors: (i) the sheer enormity of microbial taxonomic diversity in any given sample, (ii) the confounding effect of gene conservation between taxonomic groups, and (iii) the potential for relative scarcity of reads representing a pathogen, even in samples with a significant viral load. Another example in our experience was revealed when a collaborator asked why a small number of cetacean sequence reads (derived from whales and dolphins) were present within the metagenomic data from an animal slaughterhouse that was processing cattle and sheep (Hall et al., 2013). This particular study was aimed at virus discovery, and was designed to be very sensitive, using sequence similarity search parameters that allowed identification of distantly related hits to enable the detection of novel viruses with low-level homology to known sequences. However, especially in eukaryotes, genetic sequences can be highly conserved, and in this case, the homologous matches occurred due to genetic conservation between the spuriously observed cetaceans, and ruminants such as the cattle and sheep processed in the slaughterhouse.

Explaining such cases to collaborating researchers takes time and careful communication, to allay their concerns about the inherent inaccuracies of the metagenomic method. Imagine, then, how difficult it could be to allay the concerns of a patient reading a report containing a non-significant hit to smallpox, or a farmer reading a report containing a spurious hit to foot-and-mouth disease.

Considerations for Clinical Use of Metagenomics

Before incorporating metagenomics into routine clinical diagnostics, viral databases need to be vastly expanded, so that sequences can be more accurately annotated (Dutilh, 2014). Thus, efforts to map the complete viromes of humans and economically relevant animal or crop species will provide a baseline for allowing metagenomics to be applied in the clinic. Moreover, a good reference database allows novel viruses to be readily detected. Novel viruses that are observed in humans for the first time, e.g., after genomic recombination of known viruses or by zoonotic transfer from risk species like bats,

can be flagged as potentially dangerous (Temmam et al., 2014).

Moreover, general considerations need to be made, such as the time taken to generate results, how these results are reported, how performance attributes like sensitivity are assessed, and how quality assurance programs and criteria for accreditation should be developed. Indeed, recent concerns about the potential for false-positive detection of pathogens in metagenomics datasets (Naccache et al., 2014; Rosseel et al., 2014) underscore the need to develop proper quality control procedures before routine deployment of metagenomics in the clinic or diagnostic laboratory.

Large-scale deployment of diagnostic methods in clinical laboratories is facilitated by simplicity, repeatability, low costs, established quality assurance programs, and quick turn-around times for the production of results. The time and cost of processing a sample through a next-generation sequencer, albeit rapidly reducing, is still prohibitive when considering large scale diagnostic testing. Additionally, metagenomics is currently too complex for immediate release into the diagnostic laboratory. There is a lack of standardization in the laboratory methods applied, such as the choice of sequencing platform or upstream sample preparation that is used. The *ad hoc* and heterogeneous tools currently employed for the analysis of high-throughput metagenomic datasets will also need to be further streamlined and unified. Reliability values are needed to account for the conservation of identifying sequences between pathogens and non-pathogens. In addition, depending on the methodology used, even a virus present in high titer might be represented by only a few reads in a metagenomic dataset. Thus, protocols need to be optimized, and in cases where coverage of a potential pathogen is low, other diagnostic methods, such as PCR, culture, immunoassays, or electron microscopy may be necessary to confirm the presence of the pathogen.

Nucleic acid amplification technologies (NAAT) such as real-time (quantitative) PCR already offer a rapid, cheap, sensitive and specific test for application in a broad range of settings (Gray and Coupland, 2014). Existing NAAT diagnostic tests already have a high degree of utility and meet diagnostic requirements for many areas – namely those requiring the detection of specific pathogens. Ultimately, the most valuable application of metagenomics in the clinic may be to replace serial testing/multiple single-plex assays, by offering a universal metagenomics-based test. Costs and turn-around times will still need to reduce significantly, and as mentioned above, quality assurance and method standardization are areas that will need major development. The Critical Assessment of Metagenome Interpretation[3] (CAMI) is currently addressing these bioinformatic challenges in the form of a competition, by inviting

[3]http://www.cami-challenge.org/

TABLE 1 | A summary of seven major issues identified when considering the use of metagenomics as a diagnostic method, and the proposed actions that could resolve these issues.

Issue	Description of problems	Proposed actions to resolve
(1) Handling of incidental findings	Incidental detection of a pathogen that is unrelated to the investigation is possible when using metagenomics. This may be of high consequence for a patient or industry. (For industry, see point 2 below.)	Adopt protocols used for incidental findings from medical imaging studies (magnetic resonance imaging, MRI) or genome sequencing. The clinician and patient should understand the potential for incidental findings, and a plan should be in place for acting on findings as required.
(2) Agricultural/Horticultural Implications for trade	Pathogens affecting industry or trade may be detected or suspected. Even unsubstantiated reports of a high risk pathogen can have deleterious economic effects.	Independent and accredited diagnostic methods should be used to confirm the finding. Regulatory authorities should be contacted early to raise these issues.
(3) Host genome	Host genome sequence may be present in clinical metagenomic datasets. This may contravene ethical approval or legislation for handling human genome sequence (depending on jurisdiction).	Bioinformatic filtering of the host genome or restricted data access may provide some protection. Ethics committees and repositories should develop guidelines for the handling of potentially personally identifiable data in the metagenomics data.
(4) Data sharing	Deposition of metagenomic datasets from clinical samples into public databases may be problematic due to conflict with ethical, privacy, and legal concerns.	Sharing of metagenomic data is critical to the advancement of scientific understanding. However, legal and ethical constraints need to be considered and appropriate measures taken, e.g., review by ethics boards and sharing through of an appropriate data-sharing repository.
(5) Cost	Next-generation sequencing is still costly when compared to conventional diagnostic testing, especially for detecting known pathogens.	We expect that sequencing costs will continue to drop. Metagenomics is already cheaper than performing a large series of specific tests, but conventional diagnostic methods may still be preferred when searching for specific targets.
(6) Quality assurance	Currently, there are no standardized metagenomic methods: sample processing, sequencing instruments, bioinformatic analyses, and reporting of results all vary widely.	Guiding authorities will need to consider the role of metagenomics in diagnostic testing and provide protocols and quality assurance programs. For bioinformatic interpretation, the Critical Assessment of Metagenome Interpretation (CAMI) paves the way by evaluating methods.
(7) Etiology	The detection of a micro-organism in a sample does not necessarily mean it has caused the disease.	As with all diagnostic assays, prior evidence of pathogenicity or further study to determine causation (Lipkin, 2010) will be necessary to conclude that a specific organism is causing the disease.

tool developers to compete in the analysis of defined metagenomic datasets. CAMI evaluates the metagenome analysis tools and methods independently, comprehensively, and without bias. Thus, CAMI is paving the way for consistent and reliable metagenome interpretation tools to surface and receive international recognition.

Etiology

Many end users will view metagenomics as a new technology, and in terms of application outside of a research setting, it certainly is. With the advent of new technologies come high expectations. For example, given the widely held apocryphal notion that many unsolved diseases are caused by viruses (Lipkin, 2014), will such diseases of unknown etiology now be resolved? Depending on the situation, this may or may not be the case, but it is certain that there are many other possible reasons for an unresolved etiology, such as an inadequate specimen or a diagnostic test that did not include the relevant pathogen, or a toxigenic or genetic cause of the disease.

One criticism of metagenomic virus discovery projects is that the mere detection of a micro-organism in a disease-state is not sufficient to establish etiology (Canuti et al., 2014). However, this holds equally true for any other detection protocol. There are certainly high profile instances of a "virus in search of a disease," and of a "disease in search of a virus". This has become particularly apparent for syndromic diseases such as encephalitis, where orthodox testing regimes have failed to identify a cause, and for cancers or autoimmune disease of unknown etiology. Evidence for the involvement of viruses in some cases of type 1 diabetes, inflammatory bowel disease, and asthma has recently been summarized (Foxman and Iwasaki, 2011). However, it is now generally recognized and accepted that viral metagenomics is primarily suitable as a discovery and detection method, and that any claims made in regard to etiology require extensive supporting information to fulfill Koch's postulates, or variations thereof where applicable (Mokili et al., 2012).

Another criticism is that virus discovery, including viral metagenomics, is a descriptive research field that lacks hypotheses or the pursuit of knowledge about higher level biological processes. While this is a fair criticism, it should not be used to hinder or halt efforts to discover new viruses. The processes of infection, pathogenesis, or disease ecology cannot be fully understood without a basic fundamental description of the viral ecosystem, e.g., of the human virome in the case of the human body. Similarly, the development of therapeutics, vaccines, and culture methods may be informed by the discovery of new viruses. Highly divergent viral genomes may provide information about critically conserved genes and thus reveal targets for antiviral therapies, or epitopes for vaccine development. Finally, the notoriously prevalent "unknowns" in viral metagenomes can only be resolved if we face the grand challenge of mapping viral sequence space first (Dutilh, 2014).

Conclusion

Virus discovery by metagenomics is still a fresh and developing field. Huge gains have already been made in the discovery of new viral species in a wide range of host species and samples (several examples can be found in the Frontiers in Virology Research Topic "Virus discovery by metagenomics: the (im)possibilities" of which this article is part). However, the application of viral metagenomics outside of the research setting remains relatively unexplored. Importantly, the issues surrounding the use of these methods to complement or replace existing clinical diagnostic tools need to be discussed in detail. The increase in sequencing efforts to characterize the viromes of various host species will lay a foundation for further analysis of viral metagenomes by providing a reliable reference database. To facilitate this, metagenomic datasets need to be made publicly available and mined (Dutilh, 2014), but at the same time this needs to be balanced against ethical, legal, and cultural factors and potentially include filtering steps to remove sequences matching the human reference, as a safeguard for the privacy of the individual. Incidental detection or spurious reporting of viruses, especially those of high consequence for health or trade, will require special consideration. Moreover, any application of viral metagenomics outside of a research project, for example for use in the clinic, will require good communication between patients, clinicians, and scientists, including informed consent about the handling and reporting of any incidental findings. In the case of agriculture, similar communication and agreement is required between the production sector, veterinarians, and regulators. We present a summary of the seven main points of concern and our proposed actions for resolving these in **Table 1**.

History shows that care is required when delivering a new and potentially disruptive technology. No doubt, larger debate and more deeply held concerns lie ahead in the area of sequencing eukaryotic genomics, especially for the human genome, but it is worth beginning the discussion on where viral metagenomics is heading, beyond research, on the way towards application of viral metagenomics in the field and clinic.

Acknowledgments

We would like to thank Deborah Williamson and Don Bandaranayake from ESR, Christopher Weisener and Subbarao Chaganti from the University of Windsor, and Donna Gardiner from Ngā Pae o te Māramatanga for their helpful comments on the manuscript.

References

Afshinnekoo, E., Meydan, C., Chowdhury, S., Jaroudi, D., Boyer, C., Bernstein, N., et al. (2015). Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst.* (in press). doi: 10.1016/j.cels.2015.01.001

Baird, D., Greening, L., Saville-Smith, K., Thompson, L., and Tuhipa, T. (1995). "Whose genes are they anyway?" in *Proceedings of the HRC Conference on Human Genetic Information*, Wellington.

Belak, S., Karlsson, O. E., Blomstrom, A. L., Berg, M., and Granberg, F. (2013). New viruses in veterinary medicine, detected by metagenomic

approaches. *Vet. Microbiol.* 165, 95–101. doi: 10.1016/j.vetmic.2013.01.022

Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2014). GenBank. *Nucleic Acids Res.* 42, D32–D37. doi: 10.1093/nar/gkt1030

Blomstrom, A. L. (2011). Viral metagenomics as an emerging and powerful tool in veterinary medicine. *Vet. Q.* 31, 107–114. doi: 10.1080/01652176.2011.604971

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Canuti, M., Deijs, M., Jazaeri Farsani, S. M., Holwerda, M., Jebbink, M. F., De Vries, M., et al. (2014). Metagenomic analysis of a sample from a patient with respiratory tract infection reveals the presence of a gamma-papillomavirus. *Front Microbiol* 5:347. doi: 10.3389/fmicb.2014.00347

Delwart, E. (2013). A roadmap to the human virome. *PLoS Pathog* 9:e1003146. doi: 10.1371/journal.ppat.1003146

Dutilh, B. E. (2014). Metagenomic ventures into outer sequence space. *Bacteriophage* 4:e979664. doi: 10.4161/21597081.2014.979664

Dutilh, B. E., Cassman, N., Mcnair, K., Sanchez, S. E., Silva, G. G., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 4498. doi: 10.1038/ncomms5498

Foxman, E. F., and Iwasaki, A. (2011). Genome-virome interactions: examining the role of common viral infections in complex disease. *Nat. Rev. Microbiol.* 9, 254–264. doi: 10.1038/nrmicro2541

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Gonzalez, A., Pettengill, J., Vazquez-Baeza, Y., Ottesen, A., and Knight, R. (2014). "Accurate detection of pathogens in microbial samples (and avoiding the conclusion that the platypus rules the earth)," *Poster on the Platypus Conquistador Tool for Microbiome Analyses*. Available at: http://i.imgur.com/Up4mGEE.png & https://github.com/biocore/Platypus-Conquistador [accessed February 20, 2015].

Gray, J., and Coupland, L. J. (2014). The increasing application of multiplex nucleic acid detection tests to the diagnosis of syndromic infections. *Epidemiol. Infect.* 142, 1–11.

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Hall, R. J., Leblanc-Maridor, M., Wang, J., Ren, X., Moore, N. E., Brooks, C. R., et al. (2013). Metagenomic detection of viruses in aerosol samples from workers in animal slaughterhouses. *PLoS ONE* 8:e72226. doi: 10.1371/journal.pone.0072226

Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery: Microsoft Research*. Redmond, WA.

Hudson, M. (2009). Think globally, act locally: collective consent and the ethics of knowledge production. *Int. Soc. Sci. J.* 60, 125–133. doi: 10.1111/j.1468-2451.2009.01706.x

Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., et al. (2014). EBI metagenomics–a new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.* 42, D600–D606. doi: 10.1093/nar/gkt961

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. doi: 10.1371/journal.pone.0057355

Karlsson, O. E., Hansen, T., Knutsson, R., Lofstrom, C., Granberg, F., and Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecur. Bioterror.* 11(Suppl. 1), S146–S157. doi: 10.1089/bsp.2012.0077

Kehoe, M. A., Coutts, B. A., Buirchell, B. J., and Jones, R. A. (2014). Plant virology and next generation sequencing: experiences with a Potyvirus. *PLoS ONE* 9:e104580. doi: 10.1371/journal.pone.0104580

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377. doi: 10.1128/MMBR.00007-10

Lipkin, W. I. (2014). Investigating a mystery disease: tales from a viral detective. *J. Virol.* 88, 12176–12179. doi: 10.1128/JVI.00708-14

MacDiarmid, R., Rodoni, B., Melcher, U., Ochoa-Corona, F., and Roossinck, M. (2013). Biosecurity implications of new technology and discovery in plant virus research. *PLoS Pathog.* 9:e1003337. doi: 10.1371/journal.ppat.1003337

Mascalzoni, D., Dove, E. S., Rubinstein, Y., Dawkins, H. J., Kole, A., Mccormack, P., et al. (2014). International Charter of principles for sharing bio-specimens and data. *Eur. J. Hum. Genet.* doi: 10.1038/ejhg.2014.197 [Epub ahead of print].

Mason, C. (2015). *The Long Road from Data to Wisdom, and from DNA to Pathogen*. Available at: http://microbe.net/2015/02/17/the-long-road-from-data-to-wisdom-and-from-dna-to-pathogen/ [accessed February 20, 2015].

Merchant, S., Wood, D., and Salzberg, S. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675. doi: 10.7717/peerj.675

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M., and Tang, P. (2013). Metagenomics for pathogen detection in public health. *Genome Med.* 5, 81. doi: 10.1186/gm485

Ministry of Health. (2007). *Guidelines for the Use of Human Tissue for Future Unspecified Research Purposes*. Available at: http://www.health.govt.nz/system/files/documents/publications/guidelines-use-of-human-tissue-may07.pdf [accessed November 10, 2014].

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Naccache, S. N., Hackett, J. Jr., Delwart, E. L., and Chiu, C. Y. (2014). Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc. Natl. Acad. Sci. U.S.A.* 111, E976. doi: 10.1073/pnas.1317064111

New Zealand Government. (2008). Human Tissues Act [2008]. Available at: http://www.legislation.govt.nz/act/public/2008/0028/latest/DLM1154172.html [accessed November 10, 2014].

New Zealand Health Research Council. (2010). *Guidelines for Researchers on Health Research Involving Māori (Version 2)*. Available at: http://www.hrc.govt.nz/sites/default/files/Guidelines%20for%20HR%20on%20Maori-%20Jul10%20revised%20for%20Te%20Ara%20Tika%20v2%20FINAL[1].pdf [accessed November 10, 2014].

Ng, T. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., et al. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* 86, 12161–12175. doi: 10.1128/JVI.00869-12

Nuffield Council on Bioethics. (2014). *The Collection, Linking and Use of Data in Biomedical Research and Health Care: Ethical Issues*. Available at: http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf [accessed February 23, 2015].

Quinones-Mateu, M. E., Avila, S., Reyes-Teran, G., and Martinez, M. A. (2014). Deep sequencing: becoming a critical tool in clinical virology. *J. Clin. Virol.* 61, 9–19. doi: 10.1016/j.jcv.2014.06.013

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., and Van Borm, S. (2014). False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. *Transbound. Emerg. Dis.* 61, 293–299. doi: 10.1111/tbed.12251

Salter, S., Cox, M., Turek, E., Calus, S., Cookson, W., Moffatt, M., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12:87. doi: 10.1186/s12915-014-0087-z

Temmam, S., Davoust, B., Berenger, J. M., Raoult, D., and Desnues, C. (2014). Viral metagenomics on animals as a tool for the detection of zoonoses prior to human infection? *Int. J. Mol. Sci.* 15, 10377–10397. doi: 10.3390/ijms150610377

van der Hoek, L., Boom, R., Goudsmit, J., Snijders, F., and Sol, C. J. (1995). Isolation of human immunodeficiency virus type 1 (HIV-1) RNA from feces by a simple method and difference between HIV-1 subpopulations in feces and serum. *J. Clin. Microbiol.* 33, 581–588.

Woods, J. L. (2013). *Warning letter to Ms. A Wojcicki about 23andMe Personal Genome Service (PGS)*. Available at: http://www.fda.gov/ICECI/EnforcementActions/WarningLetters/2013/ucm376296.htm [accessed February 20, 2015].

Yong, E. (2015). *There's No Plague on the NYC Subway. No Platypuses Either*. Available at: http://phenomena.nationalgeographic.com/2015/02/10/theres-no-plague-on-the-nyc-subway-no-platypuses-either/ [accessed February 20, 2015].

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

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 July 2014
doi: 10.3389/fmicb.2014.00347

Metagenomic analysis of a sample from a patient with respiratory tract infection reveals the presence of a γ-papillomavirus

Marta Canuti[1†], Martin Deijs[1†], Seyed M. Jazaeri Farsani[1], Melle Holwerda[1], Maarten F. Jebbink[1], Michel de Vries[2], Saskia van Vugt[3], Curt Brugman[3], Theo Verheij[3], Christine Lammens[4], Herman Goossens[4], Katherine Loens[4], Margareta Ieven[4] and Lia van der Hoek[1]*

[1] *Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam, Academic Medical Center, University of Amsterdam, Amsterdam, Netherlands*
[2] *CBS-KNAW Fungal Biodiversity Center, Utrecht, Netherlands*
[3] *Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands*
[4] *Department of Medical Microbiology, Vaccine and Infectious Disease Institute, Universiteit Antwerpen–University Hospital Antwerp, Antwerp, Belgium*

Edited by:
Bas E. Dutilh, Radboud University Medical Center, Netherlands

Reviewed by:
Chris Sullivan, University of Texas at Austin, USA
Jan Zoll, Radboud University Medical Center, Netherlands

****Correspondence:***
Lia van der Hoek, Laboratory of Experimental Virology, Department of Medical Microbiology, Center for Infection and Immunity Amsterdam, Academic Medical Center, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, Netherlands
e-mail: c.m.vanderhoek@amc.uva.nl

† Marta Canuti and Martin Deijs have contributed equally to this work.

Previously unknown or unexpected pathogens may be responsible for that proportion of respiratory diseases in which a causative agent cannot be identified. The application of broad-spectrum, sequence independent virus discovery techniques may be useful to reduce this proportion and widen our knowledge about respiratory pathogens. Thanks to the availability of high-throughput sequencing (HTS) technology, it became today possible to detect viruses which are present at a very low load, but the clinical relevance of those viruses must be investigated. In this study we used VIDISCA-454, a restriction enzyme based virus discovery method that utilizes Roche 454 HTS system, on a nasal swab collected from a subject with respiratory complaints. A γ-papillomavirus was detected (complete genome: 7142 bp) and its role in disease was investigated. Respiratory samples collected both during the acute phase of the illness and 2 weeks after full recovery contained the virus. The patient presented antibodies directed against the virus but there was no difference between IgG levels in blood samples collected during the acute phase and 2 weeks after full recovery. We therefore concluded that the detected γ-papillomavirus is unlikely to be the causative agent of the respiratory complaints and its presence in the nose of the patient is not related to the disease. Although HTS based virus discovery techniques proved their great potential as a tool to clarify the etiology of some infectious diseases, the obtained information must be subjected to cautious interpretations. This study underlines the crucial importance of performing careful investigations on viruses identified when applying sensitive virus discovery techniques, since the mere identification of a virus and its presence in a clinical sample are not satisfactory proofs to establish a causative link with a disease.

Keywords: virus discovery, papillomavirus, respiratory tract infection, VIDISCA-454, GRACE

INTRODUCTION

Acute respiratory tract infections (ARTIs) can be caused by a wide variety of pathogens, among which viruses and bacteria are the main agents involved. Still, even with the sensitive diagnostic assays employed nowadays, a substantial amount of respiratory diseases cannot be attributed to any of the known commonly involved microorganisms (Tsolia et al., 2004; van de Pol et al., 2006; Regamey et al., 2008).

The existence of previously unknown or unexpected respiratory pathogens can be postulated as an explanation of this phenomenon. Furthermore, the pool of human respiratory viruses keeps growing because of the continuous introduction of "novel" pathogenic viruses from the animal reservoir (Neumann et al., 2009; Memish et al., 2014). These viruses need to be quickly identified and characterized, since they might represent a significant public health concern (Peiris et al., 2004). In both cases the application of broad-spectrum virus discovery techniques can be useful to detect previously unrecognized or emerging pathogens, and the introduction in the market of high-throughput sequencing (HTS) methodologies considerably improved the efficacy of such methods (Mokili et al., 2012).

One of these methods is VIDISCA-454, a restriction enzyme based virus discovery technique which was developed in our laboratory and utilizes Roche-454 as HTS method (de Vries et al., 2011, 2012). This technique provides the broad-spectrum approach needed to identify novel or unexpected viruses (Oude Munnink et al., 2014). In this study, we used VIDISCA-454 to investigate a nasal swab – that tested negative to the principal respiratory pathogens – collected from a patient with respiratory complaints to determine whether an unknown virus could be the cause of the disease.

Among the obtained sequence reads an unexpected virus was found, a papillomavirus. Papillomaviruses carry a circular double stranded DNA genome, are non-enveloped and are known for their capacity to induce warts and benign lesions of mucous membranes (condylomas; Howley and Lowy, 2007). Some human papillomaviruses (HPVs) can cause cancers, of which cervical cancer is the most notorious (Bosch et al., 1995). HPVs are not known as causative agents of respiratory infections, although some studies report the presence of different HPVs types in the respiratory tract of infants with respiratory diseases but also in healthy adults (Kurose et al., 2004; Martinelli et al., 2012; Forslund et al., 2013; Mokili et al., 2013). HPVs are also involved in recurrent respiratory papillomatosis (RRP), a disease characterized by growth of tumors around the vocal cords and in the larynx (Mammas et al., 2014). The pathogenic role of HPVs seems to be genus/type specific: cervical cancer and RRP are induced by α-HPVs (particularly types 16, 18, in cervical cancer, type 11 and 6 in RRP), while γ-HPVs are known to be causative agents of warts or skin lesions. Furthermore, several HPVs, especially from the δ- and γ-genus, can be detected on healthy normal skin (Astori et al., 1998; Antonsson et al., 2000; Li et al., 2012).

Since the high sensitivity of HTS based virus discovery methods makes it possible to identify viruses which are present in a clinical sample at a very low concentration, questions about the clinical relevance of those microorganisms might rise. The scope of this study was to determine whether a correlation could be demonstrated between the γ-HPV – identified with VIDISCA-454 – and the clinical respiratory manifestations of the patient.

MATERIALS AND METHODS

PATIENT INFORMATION

The sample was collected from a 64-year-old heavy smoking male patient (defined from now on as the index patient) during a visit to his local general practitioner where he reported symptoms of a respiratory tract infection. The onset of symptoms was 2 days before visiting his general practitioner. Symptoms included cough, phlegm production, shortness of breath, wheeze, coryza, fever, chest pain, muscle aching, and headache. During this first visit respiratory specimens (nasopharyngeal flocked swabs: NPFS, Copan) from each nostril (one in universal transport medium: UTM – the other in skimmed milk) and serum were collected (Visit 1: V1). The patient was part of the GRACE study, a randomized antibiotics placebo-controlled double-blind trial (www.grace-lrti.org), and received antibiotics (amoxicillin). A chest X-ray was performed and the patient kept a diary to monitor the course of symptoms and their severity. Fourteen days after onset of symptoms the patient fully recovered. Four weeks after the first visit the patient underwent a check up visit and again specimens from each nostril and serum were collected (Visit 2: V2). The NPFS of the first and second visit were screened for the presence of known respiratory pathogens associated with ARTI: adenovirus, respiratory syncytial virus, human metapneumovirus, influenza A and B, parainfluenza viruses 1–4, human bocavirus, human coronaviruses (229E, OC43, and NL63), human rhinovirus, polyomaviruses WU and KI, *Mycoplasma pneumoniae*, *Chlamydophila pneumoniae*, *Bordetella pertussis*, *Legionella pneumophila*, *Streptococcus pneumoniae*, and *Haemophilus* spp (Loens et al., 2012). All diagnostics remained negative.

ETHICAL APPROVAL

Ethics review committee of Cardiff and Southampton (UK) approved the study and written informed consent was provided by the participants (Loens et al., 2012).

VIRUS DISCOVERY: VIDISCA-454

VIDISCA-454 (virus discovery cDNA-AFLP, amplified fragment-length polymorphism combined with Roche 454 high-throughput sequencing) was performed with 110 μl of sample resuspended in UTM as previously described (de Vries et al., 2011). Obtained reads were compared to known sequences present in the non-redundant database using the BLAST tool (Altschul et al., 1990).

FULL GENOME SEQUENCING

The γ-papillomavirus fragments identified after performing VIDISCA-454 were used as template for primer design and PCRs were performed using the Expand Long Template PCR system (Roche). The complete genome was obtained by sequencing two large amplified fragments which were overlapping on both sides for at least 400 nucleotides. Fragments were cloned into a XL-TOPO-kit (invitrogen), and sequenced (BigDye® Terminator v1.1 Cycle Sequencing Kit, Applied Biosystems). Primer sequences used in this study are available on request.

PROTEIN EXPRESSION

The full late gene L1 (coding for the major capsid protein) was amplified with the Expand Long Template PCR system (Roche), cloned into pET100D expression vector (Invitrogen), and transformed into chemically competent Escherichia coli BL21-derived strain Rosetta 2 (Novagen). Overnight cultures of the transformed bacteria were inoculated into Luria broth medium, supplemented with 1% glucose, carbenicillin (10 μg/ml), and chloramphenicol (17.5 μg/ml). Cultures were grown to the exponential phase prior to induction with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG) for 5 h. Samples were collected every hour to monitor the L1-protein production. Attempts to purify the protein via Ni-NTA resin (Qiagen) failed, probably due to the formation of aggregates that shield the HIS-Tags; therefore the crude harvest was used to determine the reactivity of V1 and V2 serum on Western blot.

SDS PAGE AND WESTERN BLOT

Protein production on SDS-Page was visualized by coomassie staining and SDS-page gels were blotted onto PVDF (Millipore) membranes. The blots were incubated with 1:600 dilutions of either V1 or V2 serum, and with a 1:1000 dilution of a rabbit anti HIS-tag, followed by incubation with anti-human IRDye800CW (1:5000, Jackson Immunoresearch) and anti-rabbit DyLight649 (1:5000). Signals were measured via the Odyssey infrared imaging system (LI-COR). The intensity of the signals was normalized with the anti-HIS signal to correct for difference in protein concentrations.

SCREENING BY REAL TIME PCR

Real time PCRs were performed with the Platinum® qPCR-kit (Invitrogen) on an ABI PRISM 7000 sequence detection system

of Applied biosystems according to the manufacturer's protocol, using the following primers and probe: RTForward-540–570: 5′-CATACCCTAACGAAGAGGTAGCAGAC-3′, RTReverse-675–700: 5′-TGGCGGGCAACTGCCTTTATCTAG-3′, RTProbe-620–650: 5′ FAM- GGGCTGTATTCCGCCAACTGGTGAATATTGGG-ATG -TAMRA 3′.

Screening was performed in duplicate on 54 respiratory samples collected from patients with respiratory complaints involved in the GRACE study (Loens et al., 2012; 25 of which from the same city of the index patient) and only samples having detectable virus in both tests were considered positive.

RESULTS

A NPFS was collected from a 64 years old man with respiratory symptoms and the diagnostics for the most common respiratory viruses and bacteria remained negative. After performing virus discovery a total of 13,452 sequence reads were obtained and compared with all known sequences in GenBank (NCBI). Among these, 10 sequences showed significant identity to HPVs. By genome walking the entire sequence of the virus (named HPV isolate A2619) was determined: the viral genome is 7142 bp in size, circular and 99% identical to the recently described HPV-KC5 isolate (GenBank accession number: JX444073; Li et al., 2012).

CORRELATION BETWEEN HPV-A2619 AND THE RESPIRATORY DISEASE

The γ-HPVs are not known as respiratory pathogens and their role in any disease is still uncertain. Therefore, we considered it important to investigate the possible correlation between this virus and the respiratory symptoms of the patient: specific diagnostic tests were developed and applied.

Since the patient fully recovered after 14 days of illness, an increase of the specific antibody response to the responsible pathogen was expected, along with viral clearance. We examined the clearance of the virus with an HPV-A2619 strain specific real time PCR targeting the L1 gene. The virus was detected in the NPFS of both nostrils at V1, but also in the NPFS of both nostrils after the patient recovered (V2). The viral load in all samples was very low, between $2E^3$ and $2E^4$ copies/mL, and the Ct values were similar for all samples (V1-right nostril 38.93, V1-left nostril 40.60, V2-right nostril 42.90, and V2-left nostril 37.47).

Furthermore, western blot analysis was used to detect antibodies directed against the L1 capsid protein in serum samples collected at the same time as the 2 respiratory samples. A poor recognition of the protein was observed and no difference between V1 or V2 serum was noticed: the two samples had similar intensity ratios (10.78% for V1 serum; 10.76% for V2 serum; **Figure 1**).

SCREENING

Respiratory samples of 54 other patients from the GRACE study with respiratory tract infection were used for prevalence inference. Of these, 25 were from the same geographical area of the index patient and 39 from other European cities. Two patients (3.7%) were positive for HPV-A2619. In both patients also a low viral copy number was detected (between $2E^3$ and $2E^4$ copies/ml). Both patients came from the same city in the UK as the index case.

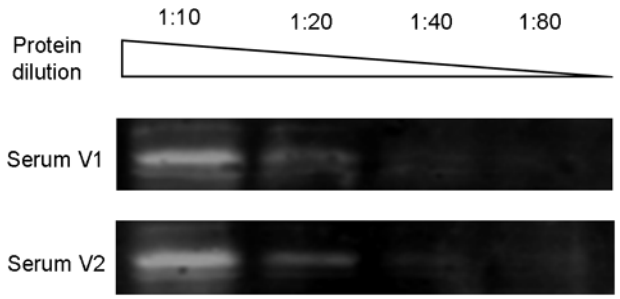


FIGURE 1 | No increase in antibody levels directed to γ-papillomavirus A2619. Visit 1 (V1) and Visit 2 (V2) sera were diluted 1:600 and incubated on Western blot containing a dilution series of L1 protein.

DISCUSSION

γ-HPVs AND RESPIRATORY INFECTIONS

Not much is known about γ-papillomaviruses (γ-HPVs). Occasionally γ-HPVs have been related to skin warts or keratotic lesions of immunosuppressed organ transplant recipients (e.g., HPV-4; Köhler et al., 2011), but these viruses can also be detected on healthy skin (Astori et al., 1998; Antonsson et al., 2000; Li et al., 2012), indicating that they might not be pathogenic in healthy individuals or only when specific conditions are fulfilled. We identified a γ-HPV in the respiratory tract of a patient with symptoms of a respiratory tract infection but, since the virus was still detectable after full recovery and no rise in antibodies was observed, we concluded that the virus was not the causative agent of the respiratory disease of this patient. Careful inspection of the chest X-ray also did not reveal any cancer-related abnormalities within the lungs; furthermore also no skin warts or abnormalities on the skin were noted. It is therefore most likely that HPV-A2619, which we describe here, is a harmless and ubiquitous γ-HPV which can be present in nostrils.

HPV-A2619 resulted 99% identical to the previously reported HPV-KC5 which was isolated from normal skin of healthy individuals in a rural area in China (Li et al., 2012), suggesting that the virus is not geographically restricted to one location, as our screening results might have indicated. Furthermore, it might be argued that the virus was introduced in the respiratory tract after contact with contaminated skin during nasal swabbing, but the fact that the virus was persistently present in each nostril at two time points suggests that virus is genuinely present in the nose, as well as on normal skin.

In literature it is described that some types of HPV can cause human papillomatosis, lesions in the respiratory tract that need to be surgically removed, and in most cases it is a recurring problem and multiple surgical treatments are needed (Mammas et al., 2014). The patient described here did not have a history of human papillomatosis, and the symptoms lasted only for 14 days, thus no indication could be found that human papillomatosis was involved. It has also been suggested that HPV, particularly type 16, can play a role in oropharyngeal cancers (D'Souza et al., 2007); however, the clinical data from this patient gave no evidences of tumor.

We therefore concluded that HPV-A2619 does not have a specific pathogenic role in a detectable disease and our results are in agreement with other studies which reported the presence of HPVs in the respiratory tract (Kurose et al., 2004; Martinelli et al., 2012; Forslund et al., 2013; Mokili et al., 2013).

SENSITIVE METAGENOMIC METHODS AND PATHOGEN DISCOVERY

Thanks to the advances in molecular biology – especially the development of next generation HTS methods – virus discovery techniques became extremely sensitive and VIDISCA-454 (the method developed in our laboratory) is not an exception (de Vries et al., 2011, 2012). The load of HPV-A2619 was determined by real time PCR, and only 200 copies per 100 μl (input in VIDISCA) were measured. This sensitivity is remarkable and might be related to the nature of the virus. HPVs carry a double stranded DNA genome and their detection by VIDISCA-454 does not depend on reverse transcription and second strand synthesis, procedures which decrease the detection potential of RNA viruses.

With state of the art HTS based virus discovery techniques and the possibility to obtain such a huge number of sequences from a clinical sample, the identification of novel viruses became very effective, and nowadays the detection of low load viruses is not uncommon since minority nucleic acids can be efficiently sequenced (Mokili et al., 2012; Cotten et al., 2014). These methods are exclusively based on random sequencing of nucleic acids and they do not provide further information beyond the recognition of viral genetic material in a clinical sample, which does not necessarily correlate with the presence of a pathogen in the sample. In fact, some of these sequences might be derived from contaminating viruses introduced during the sampling or the sample processing procedures (Naccache et al., 2013). It is therefore important to confirm the presence of the recognized viruses in the original sample and to investigate the clinical relevance of every discovery to avoid drawing false conclusions.

With the development of HTS techniques the experimental time needed to discover novel viruses has dramatically reduced. However, targeted experiments and careful evaluation of all available data are essential to assess the relevance of every finding and to determine the role of every identified virus, since – as we show here – the mere identification of a virus is not sufficient to prove its involvement in a disease. This is especially true when dealing with body parts which are contiguous with the external environment – like the respiratory tract – and therefore more subjected to the presence of bystander or non-host-specific microorganisms.

AUTHOR CONTRIBUTIONS

Marta Canuti and Martin Deijs performed the experiments, participated to study design and wrote the paper. Seyed M. Jazaeri Farsani, Melle Holwerda, Maarten F. Jebbink, and Michel de Vries performed the experiments. Saskia van Vugt, Curt Brugman, Theo Verheij, Christine Lammens, Herman Goossens, Katherine Loens, and Margareta Ieven provided the clinical samples and participated to the study design. Lia van der Hoek participated to the study design, supervised the study and helped with the manuscript preparation. All authors critically revised the manuscript and approved the final version.

ACKNOWLEDGMENTS

This study was supported by funding from the European Community's Sixth Framework Programme EC grant agreement number LSHM-CT-2005-518226 under the project GRACE and European Community's Seventh Framework Programme EC grant agreement number 223498 under the project EMPERIE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank all people involved in the GRACE Primary Care Network for the opportunity to conduct this study.

REFERENCES

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. doi: 10.1016/S0022-2836(05)80360-2

Antonsson, A., Forslund, O., Ekberg, H., Sterner, G., and Hansson, B. G. (2000). The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. *J. Virol.* 74, 11636–11641. doi: 10.1128/JVI.74.24.11636-11641.2000

Astori, G., Lavergne, D., Benton, C., Höckmayr, B., Egawa, K., Garbe, C., et al. (1998). Human papillomaviruses are commonly found in normal skin of immunocompetent hosts. *J. Invest. Dermatol.* 110, 752–755. doi: 10.1046/j.1523-1747.1998.00191.x

Bosch, F. X., Manos, M. M., Muñoz, N., Sherman, M., Jansen, A. M., Peto, J., et al. (1995). Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) Study Group. *J. Natl. Cancer Inst.* 87, 796–802. doi: 10.1093/jnci/87.11.796

Cotten, M., Oude Munnink, B., Canuti, M., Deijs, M., Watson, S. J., Kellam, P., et al. (2014). Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS ONE* 9:e93269. doi: 10.1371/journal.pone.009326

de Vries, M., Deijs, M., Canuti, M., van Schaik, B. D., Faria, N. R., van de Garde, M. D., et al. (2011). A sensitive assay for virus discovery in respiratory clinical samples. *PLoS ONE* 6:e16118. doi: 10.1371/journal.pone.0016118

de Vries, M., Oude Munnink, B. B., Deijs, M., Canuti, M., Koekkoek, S. M., Molenkamp, R., et al. (2012). Performance of VIDISCA-454 in feces-suspensions and serum. *Viruses* 4, 1328–1334. doi: 10.3390/v4081328

D'Souza, G., Kreimer, A. R., Viscidi, R., Pawlita, M., Fakhry, C., Koch, W. M., et al. (2007). Case-control study of human papillomavirus and oropharyngeal cancer. *N. Engl. J. Med.* 356, 1944–1956. doi: 10.1056/NEJMoa065497

Forslund, O., Johansson, H., Madsen, K. G., and Kofoed, K. (2013). The nasal mucosa contains a large spectrum of human papillomavirus types from the Beta-papillomavirus and Gammapapillomavirus genera. *J. Infect. Dis.* 208, 1335–1341. doi: 10.1093/infdis/jit326

Howley, P. M., and Lowy, D. R. (2007). "Papillomaviruses," in *Fields Virology*, 5th Edn. Philadelphia: Lippincott, Williams and Wilkins, 2300–2354.

Köhler, A., Gottschling, M., Manning, K., Lehmann, M. D., Schulz, E., Krüger-Corcoran, D., et al. (2011). Genomic characterization of ten novel cutaneous human papillomaviruses from keratotic lesions of immunosuppressed patients. *J. Gen. Virol.* 92, 1585–1594. doi: 10.1099/vir.0.030593-0

Kurose, K., Terai, M., Soedarsono, N., Rabello, D., Nakajima, Y., Burk, R. D., et al. (2004). Low prevalence of HPV infection and its natural history in normal oral mucosa among volunteers on Miyako Island, Japan. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod.* 98, 91–96. doi: 10.1016/s1079210404400265

Li, J., Cai, H., Xu, Z., Wang, Q., Hang, D., Shen, N., et al. (2012). Nine complete genome sequences of cutaneous human papillomavirus genotypes isolated from healthy skin of individuals living in rural He Nan province, China. *J. Virol.* 86:11936. doi: 10.1128/JVI.01988-12

Loens, K., van Loon, A. M., Coenjaerts, F., van Aarle, Y., Goossens, H., Wallace, P., et al. (2012). Performance of different mono- and multiplex nucleic acid amplification tests on a multipathogen external quality assessment panel. *J. Clin. Microbiol.* 50, 977–987. doi: 10.1128/JCM.00200-11

Mammas, I. N., Spandisos, D. A., and Sourvinos, G. (2014). Genomic diversity of human papillomaviruses (HPV) and clinical implications: an

overview in adulthood and childhood. *Infect. Genet. Evol.* 21, 220–226. doi: 10.1016/j.meegid.2013.11.002

Martinelli, M., Zappa, A., Bianchi, S., Frati, E., Colzani, D., Amendola, A., et al. (2012). Human papillomavirus (HPV) infection and genotype frequency in the oral mucosa of newborns in Milan, Italy. *Clin. Microbiol. Infect.* 18, E197–E199. doi: 10.1111/j.1469-0691.2012.03839.x

Memish, Z. A., Cotten, M., Meyer, B., Watson, S. J., Alsahafi, A. J., Rabeeah, A. A. A., et al. (2014). Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013. *Emerg. Infect. Dis.* 20, 1012–1015. doi: 10.3201/eid2006.140402

Mokili, J. L., Dutilh, B. E., Lim, Y. W., Schneider, B. S., Taylor, T., Haynes, M. R., et al. (2013). Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS ONE* 8:e58404. doi: 10.1371/journal.pone.0058404

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., et al. (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction columns. *J. Virol.* 87, 11966–11977. doi: 10.1128/JVI.02323-13

Neumann, G., Noda, T., and Kawaoka, Y. (2009). Emergence and pandemic potential of swine-origin H1N1 influenza virus. *Nature* 459, 931–939. doi: 10.1038/nature08157

Oude Munnink, B. B., Canuti, M., Deijs, M., de Vries, M., Jebbink, M. F., Rebers, S., et al. (2014). Unexplained diarrhoea in HIV-1 infected individuals. *BMC Infect. Dis.* 14:22. doi: 10.1186/1471-2334-14-22

Peiris, J. S., Guan, Y., and Yuen, K. Y. (2004). Severe acute respiratory syndrome. *Nat. Med.* 10(12 Suppl.), S88–S97. doi: 10.1038/nm1143

Regamey, N., Kaiser, L., Roiha, H. L., Deffernez, C., Kuehni, C. E., Latzin, P., et al. (2008). Viral etiology of acute respiratory infections with cough in infancy: a community-based birth cohort study. *Pediatr. Infect. Dis.* 27, 100–105. doi: 10.1097/INF.0b013e31815922c8

Tsolia, M. N., Psarras, S., Bossios, A., Audi, H., Paldanius, M., Gourgiotis, D., et al. (2004). Etiology of community-acquired pneumonia in hospitalized school-age children: evidence for high prevalence of viral infections. *Clin. Infect. Dis.* 39, 681–686. doi: 10.1086/422996

van de Pol, A. C., Wolfs, T. F., Jansen, N. J., van Loon, A. M., and Rossen, J. W. (2006). Diagnostic value of real-time polymerase chain reaction to detect viruses in young children admitted to the paediatric intensive care unit with lower respiratory tract infection. *Crit. Care* 10:R61. doi: 10.1186/cc4895

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

Received: 15 April 2014; paper pending published: 17 June 2014; accepted: 23 June 2014; published online: 08 July 2014.

Citation: Canuti M, Deijs M, Jazaeri Farsani SM, Holwerda M, Jebbink MF, de Vries M, van Vugt S, Brugman C, Verheij T, Lammens C, Goossens H, Loens K, Ieven M and van der Hoek L (2014) Metagenomic analysis of a sample from a patient with respiratory tract infection reveals the presence of a γ-papillomavirus. Front. Microbiol. ***5****:347. doi: 10.3389/fmicb.2014.00347*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 February 2015
doi: 10.3389/fmicb.2015.00097

A metagenomic approach to characterize temperate bacteriophage populations from Cystic Fibrosis and non-Cystic Fibrosis bronchiectasis patients

Mohammad A. Tariq[1†], Francesca L. C. Everest[1†], Lauren A. Cowley[2], Anthony De Soyza[3,4], Giles S. Holt[1], Simon H. Bridge[1], Audrey Perry[3], John D. Perry[3], Stephen J. Bourke[5], Stephen P. Cummings[1], Clare V. Lanyon[1], Jeremy J. Barr[6] and Darren L. Smith[1]*

[1] Faculty of Health and Life Sciences, University of Northumbria at Newcastle, Newcastle Upon Tyne, UK
[2] Public Health England, London, UK
[3] Freeman Hospital, Newcastle Upon Tyne, UK
[4] Institute of Cellular Medicine, Newcastle University, Newcastle Upon Tyne, UK
[5] Royal Victoria Infirmary, Newcastle Upon Tyne, UK
[6] Department of Biology, San Diego State University, San Diego, CA, USA

Edited by:
Alejandro Reyes, Universidad de los Andes, Colombia

Reviewed by:
Steven Ripp, University of Tennessee, USA
Samuel Schwartz Minot, Signature Science, LLC, USA

***Correspondence:**
Darren L. Smith, Applied Sciences, University of Northumbria, Ellison Building, EBD222, Newcastle Upon Tyne NE1 8ST, UK
e-mail: darren.smith@northumbria.ac.uk

† *Co-lead author's: Mohammad A. Tariq and Francesca L. C. Everest.*

Pseudomonas aeruginosa (Pa), normally a soil commensal, is an important opportunistic pathogen in Cystic Fibrosis (CF) and non-Cystic Fibrosis Bronchiectasis (nCFBR). Persistent infection correlates with accelerated decline in lung function and early mortality. The horizontal transfer of DNA by temperate bacteriophages can add gene function and selective advantages to their bacterial host within the constrained environment of the lower lung. In this study, we chemically induce temperate bacteriophages from clonal cultures of Pa and identify their mixed viral communities employing metagenomic approaches. We compared 92 temperate phage metagenomes stratified from these clinical backgrounds (47 CF and 45 nCFBR Pa isolates) using MG-RAST and GeneWise2. KEGG analysis shows the complexity of temperate phage accessory gene carriage increases with duration and severity of the disease. Furthermore, we identify the presence of Ig-like motifs within phage structural genes linked to bacterial adhesion and carbohydrate binding including Big_2, He_Pig, and Fn3. This study provides the first clinical support to the proposed bacteriophage adherence to mucus (BAM) model and the evolution of phages interacting at these mucosal surfaces over time.

Keywords: *Pseudomonas aeruginosa*, temperate bacteriophage, cystic fibrosis, non-Cystic fibrosis bronchiectasis, mixed phage populations

INTRODUCTION

Chronic respiratory diseases are associated with about 4 million deaths globally per annum (WHO[1]). Cystic Fibrosis (CF) is a rare but well documented inherited chronic respiratory disease that is characterized by chronic bacterial infections of the lung (Mall and Boucher, 2014). Non-Cystic Fibrosis bronchiectasis (nCFBR) is usually associated with an older population; it is an abnormal and irreversible dilation of the lower bronchi. Unifying features between CF and nCFBR are the propensity for *Pseudomonas aeruginosa* (Pa) to be an opportunistic pathogen and abnormal mucus retention in the lower respiratory tract.

Pa is challenging to study at the genome level due to the presence of multiple genomic islands and a burgeoning accessory genome that may well correlate with its opportunistic nature. Opportunistic bacteria colonize the lungs of patients with chronic respiratory disease and utilize the nutrient rich mucus lining of the lower airways allowing for bacterial replication and evolution to occur in an often deteriorating microenvironment (Nelson et al., 2010; Hauser et al., 2011; Rudkjobing et al., 2011). Other bacteria that are commonly isolated from chronically infected lung include; *Staphylococcus aureus, Haemophilus. influenzae, Stenotrophomonas maltophilia, Achromobacter xylosoxidans,* whilst *Burkholderia cepacia* complex and Pa are descriptive of the CF lung and are linked to poor clinical outcomes including lowered lung function (Lipuma, 2010). Multiple phages have been previously identified in Pa isolated from the CF lungs (Winstanley et al., 2009).

Bacteriophages can be either classed as lytic or temperate. Lytic phages upon entry into their host bacterium rapidly propagate leading to cell lysis. Importantly, lytic phages do not become integrated into the bacterial chromosome; this is in contrast to temperate phages which upon entry into the cell integrate into the host genome as a prophage. Temperate phages infect their bacterial host and lay dormant within the bacterial host chromosome until they are induced from their host, forming an infective phage particle. It has been noted that phages typically outnumber bacteria by a factor of 10 (Fineran et al., 2009). Previous metagenomic

[1]WHO. Available online at: http://www.who.int/genomics/public/geneticdiseases/en/index2.html#CF [Accessed February 20, 2014].

studies focusing on viruses have identified novel patterns associated with evolution and novel viral particles (Kristensen et al., 2010). In this study, temperate bacteriophages were induced from their bacterial host using Norfloxacin (Matsushiro et al., 1999).

The major focus of numerous genome studies is determination of the core phage genome architecture. This study uses a metagenomic approach to elucidate the depth, function and complexity of phages evolving in a constrained environment of the lower lung. Conventional genome assembly tools try to compile mixed communities into single phages as they try to match and overlay similarity of sequence composition. Here we employ Metagenomics Rapid Annotations based on Subsystem Technology (MG-RAST) to overcome the need to assemble single phages and focus on the accessory genomes functionality. Another advantage of using MG-RAST is to generate Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways; these allow for analysis of gene functionality via linking genetic information with higher order functional information (Kanehisa and Goto, 2000).

Here we focused on lysogenic phages, as they form an intrinsic part of the adaptation and evolution of bacteria (Bankevich et al., 2012). Temperate phages have been shown to carry a range of genes that can alter the fitness or pathogenicity of a bacterium. An example would be the ability to encode functional toxins in their host bacterium which in turn can increase the severity of disease and may influence its progression (Beddoe et al., 2010; Boyd et al., 2012; Dubreuil, 2012). Our focus was on phage accessory genes, which link temperate phages to bacterial adaptation and evolution in chronic lung infections. Phage accessory genes are understudied as they are small with no offered function; importantly these genes are usually shared between phages suggesting a conserved role in their biology or for subversion of their host (Smith et al., 2012). Here we compare clinical data and the complexity of phage-encoded accessory gene function to link to the pathophysiology of the chronic lungs in patients with CF and nCFBR. Metagenomic studies are beneficial for studying mixed viral communities as they utilize culture independent methods allowing for the observation of viral communities that lack a known propagating host and therefore, can be underrepresented in some studies. This is also a problem in bacterial studies with the inability to culture all strains in the laboratory so increasing the need for direct DNA sequencing methodologies that limit culture bias (Hugenholtz et al., 1998). Further complexity is added to mixed population genome assemblies as bacteria and viruses carry homologous genes with conserved order which can make them harder to separate bioinformatically (Drancourt et al., 2000; Boudewijns et al., 2006). Metagenomics can further be utilized to investigate pan-functionality in a sample compared to more traditional taxonomic approaches (Tringe et al., 2005), making it an ideal tool for *de novo* studies as it negates the need for previous knowledge of the sample (Roux et al., 2014).

Mucus forms the first line of defense for protection against pathogenic infection in the lung forming a physical barrier between the center of the airways and the underlying epithelial cells (Hansson, 2012; Barr et al., 2013). Mucus is composed mostly of mucin, a host produced glycoprotein but other macromolecules are also known to be present (Kim and Ho, 2010). The Bacteriophage Adherence to Mucus (BAM) model proposes that lytic phages adhere to carbohydrate residues present within the mucus layer, and provide a layer of immunity to incoming bacteria (Barr et al., 2013). The BAM model is mediated by structurally displayed carbohydrate-adherence domains, such as the immunoglobulin (Ig)–like domain present on the capsid of phage T4 (Hoc). Structural proteins with associated Ig-like domains have been found in approximately 25% of the sequenced dsDNA phages, demonstrating their ubiquity and their potential importance in aiding phage survival (Fraser et al., 2007). Such structural domains have been seen to be indispensable for phage propagation in laboratory settings as they mediate interactions between the phage and its host cell (McMahon et al., 2005; Fraser et al., 2007).

Here we investigate whether the temperate phages isolated from the mucus rich environments of CF and nCFBR patients' lungs support the BAM model and its clinical relevance. We utilize the BAM model to propose a different strategy for phages to disseminate across their host These observations led to the hypothesis that these domains may aid in both the adsorption of phages to their bacterial hosts and the mucus layer under certain environmental conditions (Fraser et al., 2007). We hypothesized that temperate phages may use the mucus barrier as a way of infecting incoming bacteria which may drive gene exchange and add another level of adaptation and evolution. It also may be a way of increasing genetic heterogeneity in a population of bacteria in late stage chronic lung infections that are traditionally thought to be somewhat clonal. We compare the inducible temperate phages of Pa found in the lungs of patients with CF and nCFBR using a metagenomic approach and study the complexity of putative functional traits the phage accrue through their continual adaptation and evolution in the chronic lower lung.

MATERIALS AND METHODS

BACTERIAL ISOLATES AND MEDIA

The Pa isolates in this study originate from clinical isolates collected at the Freeman Hospital and the Royal Victoria Infirmary; Newcastle Upon Tyne Hospital Trust, UK (10 pediatric CF isolates, 37 adult Cystic Fibrosis (CF) isolates, 17 $<$ 10 year clinical diagnosis nCFBR isolates and 28 $>$ 10 year nCFBR isolates). All bacterial cultures were propagated in Luria Broth (LB) media (Sigma Aldrich, Gillingham, UK), $CaCl_2$ was added to Soft Agar [0.4% high clarity agar (Lab M Limited, Heywood, UK) and 0.01 M $CaCl_2$ (Sigma Aldrich)] to promote phage adsorption; the cultures were incubated at 37°C for 18 h (+200 rpm if liquid culture). Full ethical approval has been given for this work (REC reference: 12/NE/0248).

PROPHAGE INDUCTION AND RE-INFECTION

Lysogenic bacteriophages were chemically induced from bacterial isolates. In brief, overnight cultures were sub cultured 0.2% (v/v) (10 mL LB Broth, 0.01 M $CaCl_2$). The phages were induced by stressing the bacterium with fluoroquinolone antibiotic, Norfloxacin (1 $\mu g.mL^{-1}$) (Sigma Aldrich) for 1 h (37°C, 200 rpm). The culture containing norfloxacin was diluted (1:10) to limit the cytotoxic effect of the drug, so allowing for the cascade of phage induction to occur (Matsushiro et al., 1999).

Phage lysates were filtered through a 0.22 μM filter (Scientific Laboratory Supplies, Hessle, UK) and stored at 4°C for <1 week. Phage lysates were also utilized to identify whether a phage from the lysate had the ability to re-infect the originating bacterial host. The lysates were spotted (10 μl) in dilution onto a lawn of originating host Pa cultured in 0.4% (w/v) agar + Luria Broth, overlaid on LB agar. Dilution identified plaques over possible pyocin production.

PHAGE DNA ISOLATION

Bacterial chromosomal DNA was attenuated using 1 μL of TURBO DNAse and 1 μL of RNAse Cocktail (Life Technologies Limited), prior to incubation at 37°C for 30 min followed with heat inactivation at 65°C and 0.5 M EDTA. NORGEN Phage DNA Isolation Kits (Geneflow Limited, Lichfield, UK) were used to purify viral DNA, in accordance with manufacturer's protocol. The NORGEN phage DNA isolation kit was chosen due to its optimal yield of phage DNA compared to the QIAGEN QIAmp MinElute Viral Spin Kit, Chelex extraction, and PEG8000 purification (Sambrook et al., 1989) whilst limiting bacterial chromosome contamination. A low level of bacterial chromosomal contamination was determined by PCR for the 16S rRNA gene but it was negated bioinformatically using the Khmer toolkit (Muyzer et al., 1993).

NEXT GENERATION DNA SEQUENCING

The Illumina Nextera XT (Illumina, Saffron Waldon, UK) library preparation kit was used to prepare and multiplex the isolated phage DNA for next generation sequencing on the Ilumina MiSeq platform. A 2 × 250 cycle V2 kit was used for the loading and running of the sample. The DNA samples were diluted to 0.2 ng/μL (Qubit 2.0 DS HS DNA Kit [Life Technologies Limited]) prior to normalization and pooling. Paired end sequencing reads where provided as FASTQ files (NU-OMICS, Northumbria University at Newcastle, UK) and subject to downstream analysis.

BIOINFORMATIC TOOLS

RANDOMIZING DNA READS USING VELVET V1.2.10

Velvet *de novo* genome assembler package shuffleseq.pl was used to randomly shuffle the FASTQ sequences to limit bias. The shuffled sequence output file was directly pipelined into the Khmer toolkit. Command line script can be found in Supplementary Material S6.

KHMER TOOLKIT

Khmer uses a probabilistic Bloom filter to separate out k-mer abundance and to group accordingly. This is achieved by creating hash tables that store specific k-mers and their counts using the default settings. The toolkit was utilized to remove very low-level bacterial contamination from the viral sequence data. The Khmer histogram clusters low abundance data and poor sequence data that would be linked to any residual bacterial chromosomal DNA where we can use the defining python script to select abundant viral k-mer sequencing data (Brown et al., 2012). Command line script can be found in Supplementary Material S6. For each individual sequence file we assessed the out.hist files graphed in excel and manually remove the error k-mer peak. These output files then can be pipelined to MG-RAST once the data is renamed using "Rename Sequences" v 0.0.11 (Blankenberg et al., 2010).

RENAMING FASTA FILES

Following the separation of the raw data into separate sequence files via Khmer, the files were converted from FASTQ to FASTA using a python script (Khmer package). Sequences were renamed numerically using "Rename Sequences" v 0.0.11 (Blankenberg et al., 2010). These 92 viromes where then uploaded to MG-RAST. From the 92 files and sequence cleanup 10 files had under the threshold of data allowed for uploading onto MG-RAST. However, we still assembled each of these samples using the three assemblers and used them to search for any putative Ig-like domains.

MG RAST

The KEGG generator function was used in order to show possible differences in the biochemical pathways between the phage isolated from the 2 clinical origins and stratified in the methods section. The maps were generated with the MG-RAST default setting: 60% sequence similarity of 15 amino acids. The hierarchical classification system tab was used in order to generate Principle Component Analysis (PCoA) on the samples relating to their functionality, the data was normalized and drawn according to the Minkowski distance with MG-RAST default settings.

THREE-WAY GENOME ASSEMBLY COMPARISON

Before we could search the sequence data for putative carbohydrate binding motifs using GeneWise2 we perform a three-way assembly study using SPAdes v 3.1.0, Velvet optimizer v 2.2.5 and IDBA-UD v 1.1.1. An overview of the assemblies is provided in Supplementary S5 that details each assembly comparison showing N50 scores, number of contiguous sequences derived and the largest contig size. **Figure 5A** shows the ability to detect Ig-like domains compared between the assemblers.

HMM/PFAM DATABASE SEARCHES USING GENEWISE2 V2.2.0

GeneWise2 was used to search a database of 92 sequence files against the Pfam database of 40 amino acid based HMMs as shown in **Table 1**. HMMER v 3 was used to revert the HMMs from version 3 to version 2, so GeneWise2 could recognize these files. GeneWise2 algorithms 6:23 and 21:93 were used and comparisons drawn with Jalview v 2 (Waterhouse et al., 2009). The gene locations of the resulting positive results were compared to the putative ORF associated with the GeneWise2 identification.

RESULTS

Putative mixed viral communities were induced from a cross-sectional panel of 92 *Pseudomonas aeruginosa* (Pa) bacterial isolates, 47 isolated from CF patients and 45 nCFBR stratified further by patient clinical information detailed in the methods. Phage DNA was isolated away from bacterial chromosome and any low level remaining bacterial DNA was removed bioinformatically using Khmer toolkit as described both in the methods and Supplementary Material.

Table 1 | The 40 Pfam databases used in GeneWise2 and adapted from Fraser et al. (2006).

SCOP superfamily	PFAM name	Accession number
Ig	V-set	PF07686
	I-set	PF07679
	C2-set	PF05790
	C1-set	PF07654
	Ig	PF00047
	Ig_2	PF13895
	ICAM_N	PF03921
E-SET	Alpha_amylase_N	PF02903
	arrestin_N	PF00339
	arrestin_C	PF02752
	CelD_N	PF02927
	peptidaseC25	PF03785
	TIG	PF01833
	RHD	PF00554
	DUF291	PF03442
	Filamin	PF00630
	He_Pig	PF05345
Fibronectin type 3	FN3	PF00041
	tissue_fac	PF01108
	lep_receptor_Ig	PF06328
PKD	PKD	PF00801
	PPC	PF04151
	HYR	PF02494
β-Galactosidase/β -Glucuronidase	Glycol_hydro_2	PF00703
Cu, Zn Superoxide dismutase-like	Sod_Cu	PF00080
PapD-like	Pili_assembly_C	PF02753
	pili_assembly_N	PF00345
Invasin/intimin cell-adhesion fragments	Big_1	PF02369
	Big_2	PF02368
	Big_3	PF07523
	Big_4	PF07532
Clathirin adaptor appendage domain	Alpha_adaptin_C2	PF02883
Transglutaminase N-terminal domain	Transglut_N	PF00868
Cadherin-Like	Cadherin domain	PF00028
Actinoxanthin-like	Neocarzinostatin family	PF00960
CBD9-like	Domain of unknown function	PF06452
laminA/C globular tail domain	Intermediate filament tail domain	PF00932
Other Ig-like	C type Lectin	PF00059
	BACON	PF13004
	MucBP	PF06458

They were individually selected as they encode putative carbohydrate binding domains.

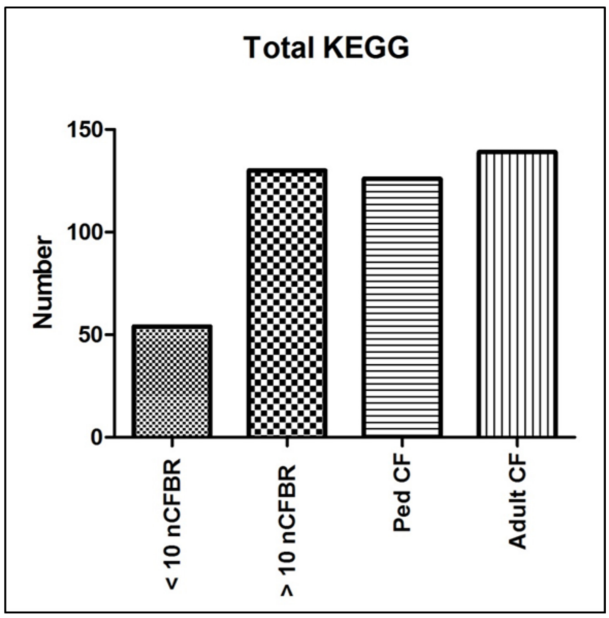


FIGURE 1 | Total KEGG putative function identifications derived for phage lysates. Each putative functional pathway is represented as a single identification regardless of the amount of times that the putative function may have been conferred in each of the clinical subgroups. The highest rates of identification are seen in the lysates that originate from adult CF patients whilst the lowest rates of identification are seen in the lysates from <10 year clinical diagnosis nCFBR patients.

MG-RAST DERIVED KEGG PATHWAY ANALYSIS

Eighty two from ninety two isolates were analyzed through MG-RAST and KEGG analysis due to 10 of the sequencing files having insufficient sequence data for analysis. The pathways that possibly confer putative function are shown in the Supplementary Data alongside the raw data generated from each KEGG pathway (S1 and S2 respectively). Using KEGG pathway analysis through MG-RAST we defined the "incidence" of DNA sequences with shared similarity with a known metabolic pathway stored in the KEGG database. An overview of the number of reads per sample is shown in Supplementary S3 whilst the KEGG derived EC values downloaded from the KEGG generator tab on MG RAST are shown in Supplementary Table S4.

Using KEGG pathway analysis **Figure 1** identifies the presence of different metabolic pathways that are stored in this database. We stratify this further using the clinical information detailed in the methods. We identify an increase in KEGG derived pathways that link to the duration of clinical disease in adult CF and >10 years nCFBR. However, nCFBR (<10 year clinical diagnosis) lysates have the lowest number of KEGG pathway incidences. There is an overall increase in the number of identifications as both diseases progress. This may show phage adaptation and accrual of genes that possibly aid the fitness of the bacterial host within this environment. **Figure 2** shows bar graphs demonstrating the incidence of KEGG identifications that have been stratified both by clinical etiology but also by the defining zones of the KEGG atlas. The KEGG pathway analysis identifies gene regions and links to functional pathways relating to metabolism

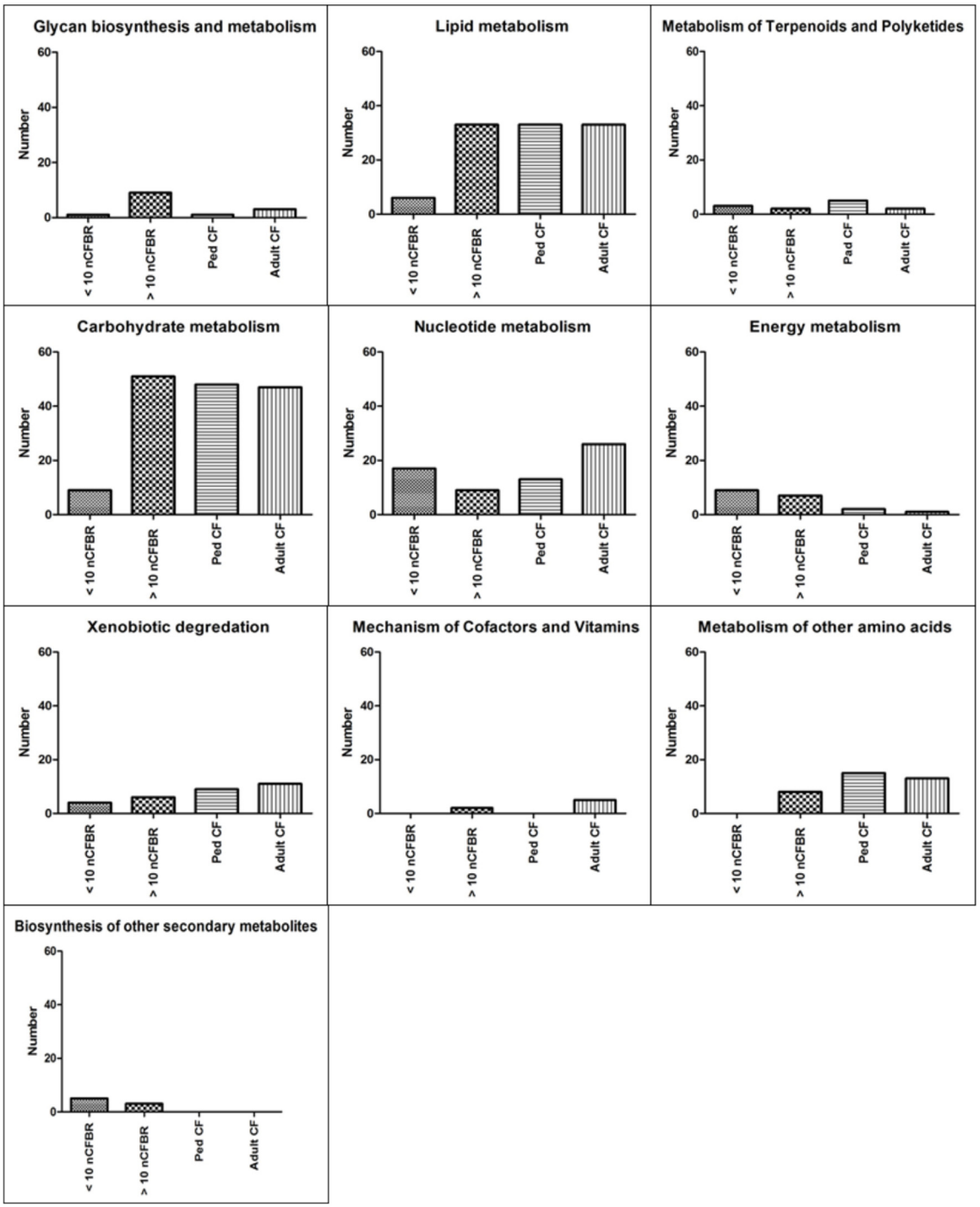


FIGURE 2 | Incidence of amino acid similarity to KEGG functional pathways generated by MG-RAST. These data are further stratified sub-etiology to; pediatric CF (ped CF), adult CF, < 10 year clinical diagnosis nCFBR and >10 year clinical diagnosis nCFBR. Each putative functional pathway is represented as a single identification regardless of the amount of times that the putative function may have been conferred in each of the clinical subgroups. These KEGG pathways can be used to confer possible functionality and difference between the disease states. A clear increase in the number of KEGG pathway identifications is linked to patient age regardless of disease etiology ("Glycan biosynthesis and metabolism," "Xenobiotic degradation," and "Mechanism of Cofactors and Vitamins"). This pattern is reversed in some panels with a reduction in the number of identifications decreasing as the disease progresses regardless of etiology ("Metabolism of Terpenoids and Polyketides" and "Energy Metabolism"). When looking Nucleotide metabolism there is a decrease in the amount of identifications between <10 year clinical diagnosis nCFBR and >10 year clinical diagnosis nCFBR but an increase between ped CF and adult CF patients. In two panels ("Lipid Metabolism" and "Metabolism of other amino acids") there is also an increase in the number of identifications seen as nCFBR patients disease state progresses but a decrease in the number of hits as a CF lung deteriorates. When looking at the final panel ("Biosynthesis of other secondary metabolites") no similarities are detected for the CF and nCFBR patients regardless of progression of disease.

and signaling. Increases in functionality that can be linked to disease progression include; glycan biosynthesis and metabolism; xenobiotic degradation; mechanisms of cofactors and vitamins. An overall decrease in identifications was identified in pathways relating to; metabolism of terpenoids and polyketides, and energy metabolism. However, when looking at some areas of the KEGG atlas there are differences between the CF and nCFBR lysates. Lipid, carbohydrate and amino acid metabolism increases as nCFBR progresses where the opposite is apparent for CF patients. The reverse of this trend is observed in hits relating to nucleotide metabolism, with a decrease in hits in nCFBR patients compared to an increase in CF patients. It is noteworthy that when comparing identifications relating to the biosynthesis of secondary metabolites it is clear that there are no hits in a CF phages background.

MG-RAST derived PCoA

Principal component analysis (PCoA) generated by pan differences in functionality (**Figure 3**) illustrates that the phage samples show unrelated functionality and that there is no observable link between the phage functionality on the PCoA. This suggests that even though differences in function are seen when looking at the KEGG atlas (**Figures 1**, **2**); these differences are not discrete enough to lead to separation on the PCoA plot and thus show that not all phage carry all traits.

Phages ability to Superinfect Pa

A large number of phages that were chemically induced in a previous study from the 94 isolates have the capability of re-infecting their originating bacterial host (**Figure 4**). Stratification according to disease origin has been shown in **Figure 4A** where 49% of nCFBR related phages had the ability to re-infect the originating host compared to 66% of the CF Pa induced phages. In **Figure 4B** further stratification illustrates that 16% of pediatric CF Pa phages within these samples and 84% of the adult samples have the ability to re-infect. This trend is also seen in nCFBR patients with 30% <10 years clinical diagnosis and 70% >10 years clinical diagnosis able to re-infect their originating host.

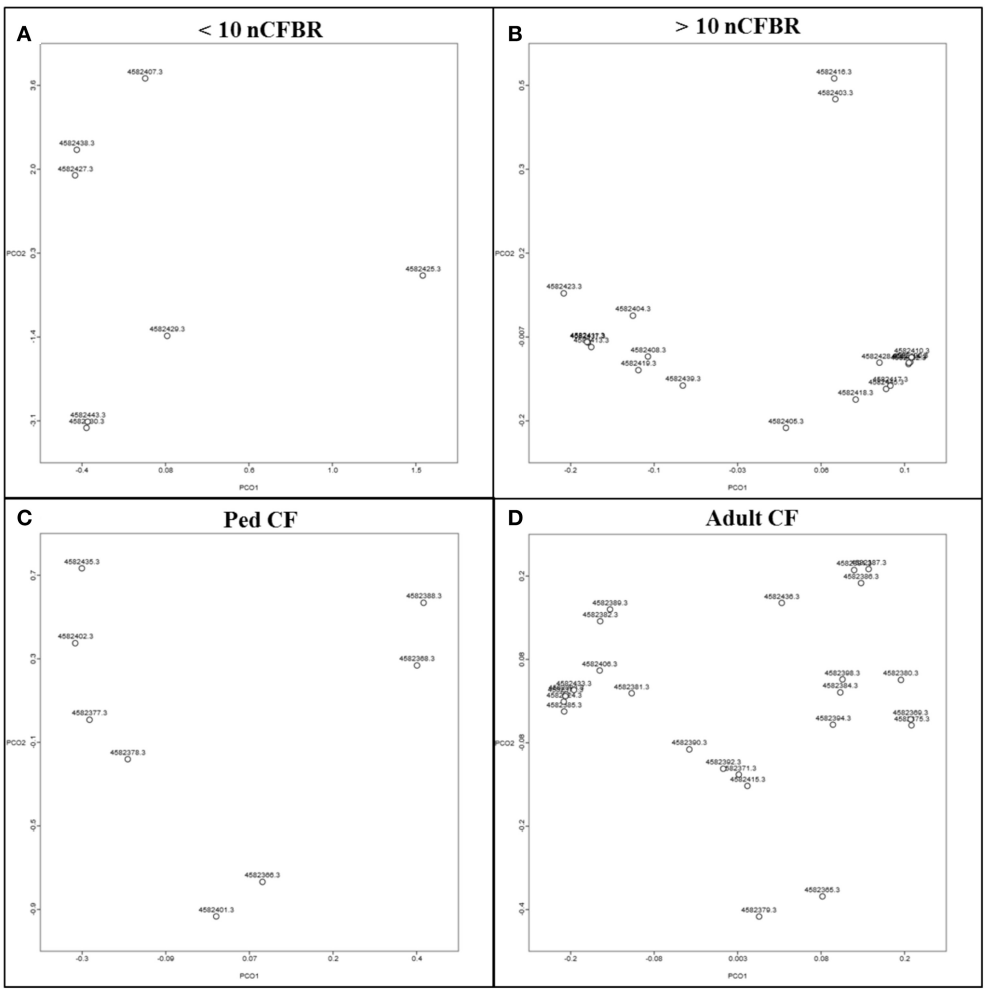


FIGURE 3 | Principle component analysis drawn in MG-RAST according to COG and the Minkowski distance. They indicate that the phage lysates postKhmer analysis have no relatedness to the presence of a Big_2 domain in each sample. **(A)** Represents <10 year clinical diagnosis nCFBR samples, **(B)** represents >10 year clinical diagnosis nCFBR samples, Ped CF samples are shown in **(C)** and the adult CF samples are shown in **(D)**.

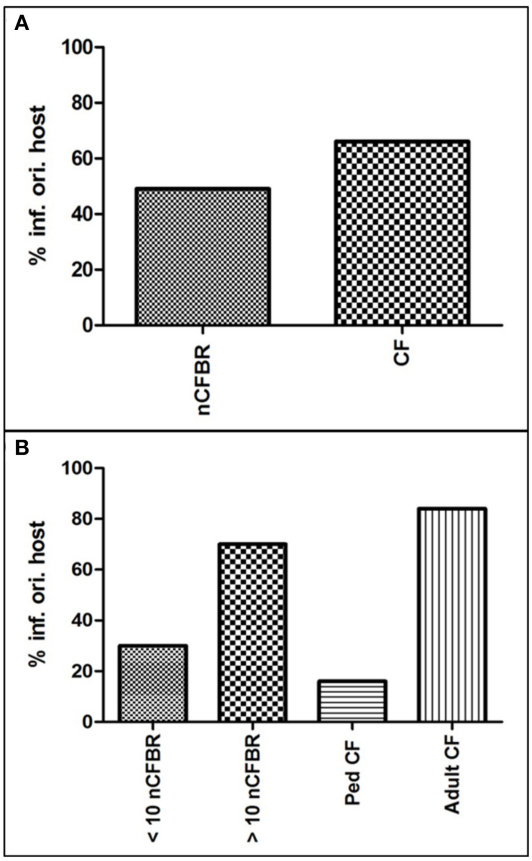


FIGURE 4 | Graphical representation of the percentage of phages that are capable of re infecting their originating host (% inf. ori. host = % infection of originating host). (A) Shows the percentage of phage lysates from both nCFBR and CF, with CF phage having greater ability to infect their originating host. **(B)** shows the stratification of these rates shown in **(A)**, illustrating that this trait is acquired over time in phages isolated from both clinical etiologies.

Presence of Ig-like domain Big_2

The frequency of Ig-like binding domains is described in **Figure 5**. It illustrates that the frequency of Big_2 domains increases alongside the longevity of disease. This has been calculated as a percentage of the total number of phages for each of the clinical strata, which have been seen to contain a Big_2 domain (CF: pediatric CF phages 30%, adult CF phages 40%; nCFBR: <10 years clinical diagnosis 6%, >10 years clinical diagnosis 32%). This work is unique as these domains have not been observed on phage genomes to this level previously, especially on phages with known clinical origins.

Identification of Ig-like binding domains

Big_2 domains were only identified with both GeneWise2 6:23 and 21:93 algorithm and these were seen in 28 of the 92 mixed phage samples. All bit similarity scores were above the recommended cut-off of 25, were considered significant for gene prediction (Fraser et al., 2006). The domain architecture found in these samples is shown in **Figure 6A**. It was seen that the Big_2

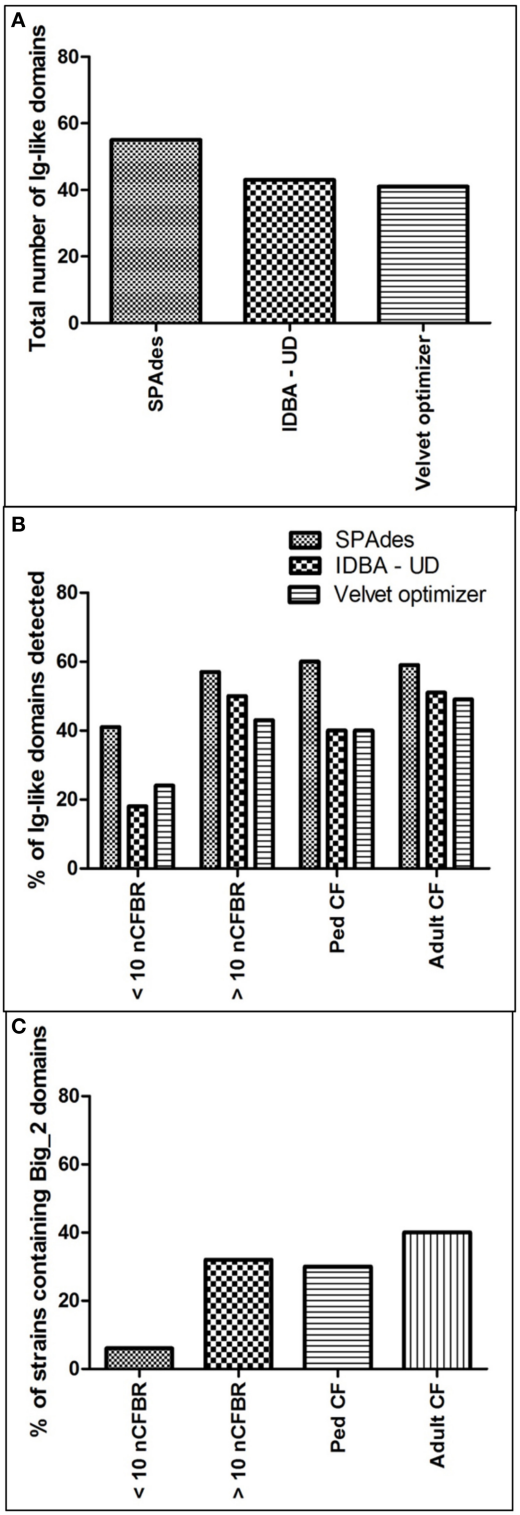


FIGURE 5 | (A) Describes the total number of Ig-like domains identified using Genewise2 subsequent to assembly by SPAdes, IBDA-UD, and Velvet Optimizer. **(B)** Offers the percentage of the total number of Pa isolates from the 4 clinical groups with one or more Ig-like domain. **(C)** Focuses on the percentage incidence of the Big_2 domains which attracted the highest bit scores of detection using GeneWise2.

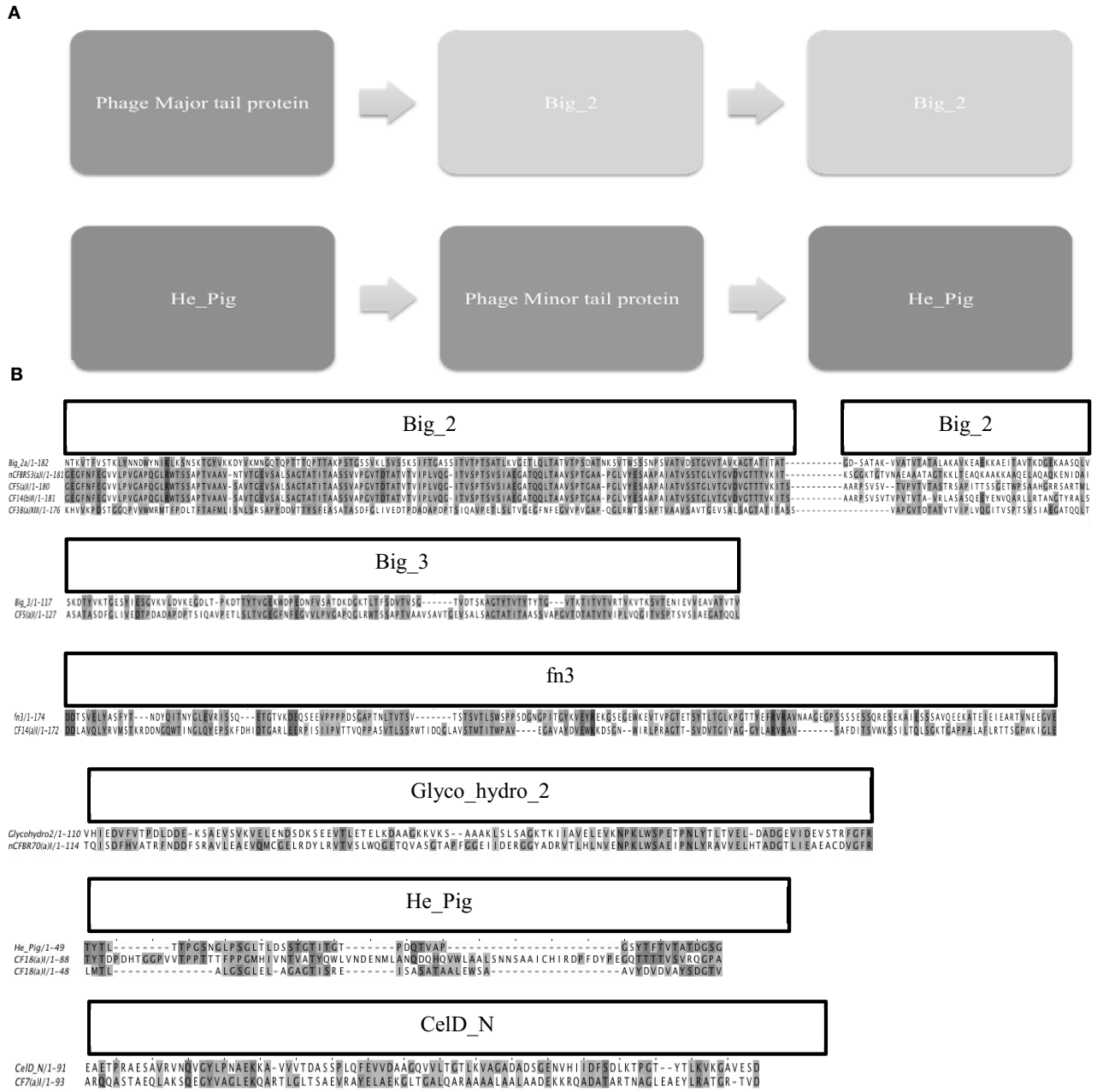


FIGURE 6 | (A) Shows the domain architecture of the double Big_2 domain that was identified in a subset of the phage lysates from the clinical samples, the second image is the proposed domain architecture for He_Pig. The He_Pig domains however, span over different genes adjacent to the Phage minor tail protein. **(B)** Shows the sequence alignments of the Ig-like domains; Big_2, Big_3, Fn3, Glyco_hydro_2, He_Pig and CelD_N found in phage proteins. The sequence alignment of Big_2 is represented along with 4 variations that were seen in this study; all the samples have a high similarity to the Big_2 domain represented on the top line. Big_3 and Fn3 both have a lower sequence identity to the known sequence but this alignment was conserved between all the phages that were identified to contain these domains. The Glyco_hydro_2 domain was seen in only one phage sample (> 10 clinical diagnosis nCFBR isolate). The He_Pig double domain was seen in 42 of the phage samples, 70% of which have the amino acid alignment shown above.

domains were found in duplicate and associated with the major tail protein. It was also seen using GeneWise2 algorithm 21:93, where a double He_Pig domain presented in a large subset of the Pa phages as a double motif in putative structural genes flanking a putative minor phage tail protein. The He_Pig double domain was seen in 42 of the phage samples, over 70% of which have the amino acid alignment shown in **Figure 6B**. All the other hits to the He_Pig domain showed poor bit scores and were thus not aligned in **Figure 6B**.

When the GeneWise2 21:93 algorithm was used, it was seen that a Big_3 domain was present overlapping (by 89 amino acids) the first Big_2 domain which may link to a frameshift that

has been described in other Ig-like binding motif architecture (Fraser et al., 2006). This Big_2, Big_3 combination trend was seen in all of the samples, except for four where the algorithm failed to report any significant similarity. We also identified another domain in 10 of the samples called Fn3. The domain alignments for Big_3, Fn3, Big_2, glycol_hydro_2, CelD_N, and He_Pig are shown in **Figure 6B**. The Glyco_hydro_2 domain was identified but not associated with a putative structural protein. Upon utilizing a BLASTn search the Fn3 domain was located in a phage tail assembly protein. Similarly the ORF for Big_2 was also identified in a structural major tail protein gene, this time the major tail 2 protein. Other Ig-like domains have been identified but they all overlapped the first Big_2 domain including; PPC, Peptidase_C25_C, Big_3, and Big_4 all had too low bit scores.

DISCUSSION

This study reports the first use of metagenomic approaches to identify the inducible temperate bacteriophages isolated from single clonal cultures of *Pseudomonas aeruginosa* (Pa) colonizing the lungs of patients with CF and nCFBR. These data were further compared to the clinical information provided for each patient sub-group. A noteworthy observation is the evolution in complexity of the viral population linking to possible functions that aid bacterial fitness and therefore, viral sustainability in chronic lung infections. Key findings are the high frequency of Ig–like motifs found in the temperate phages of Pa with known clinical origin and also the increase in complexity of putative phage function when adapting to the progressive disease state of the lower lung.

When analyzing the metagenomic results generated via KEGG analysis through MG-RAST, it is clear to see an increase in complexity of function and possibly the level of adaptation occurring between the phage in accordance with either patient age or colonization time with Pa (**Figures 1**, **2**). Using KEGG we also observed specific functional trends encoded by temperate phages that are both similar and disparate between the disease states. An increase is seen in the number of identifications for glycan biosynthesis and metabolism; this is possibly due to these functions being associated with cell wall synthesis in the bacteria and possibly associated with inflammation in the lung. The CF lung also contains a large amount of human produced mucins which are covered in glycans, so it is possible that the phage are transporting functions relating to glycan biosynthesis in order to increase the degradation of these mucin's and thus promote bacterial growth in the CF lung. Phages therefore, may utilize these functions in order to survive, by driving further inflammation in the lower lung environment offering preferential selection for the bacterium. When focusing on certain subsections such as "carbohydrate metabolism" and "nucleotide metabolism," apparent differences are seen between some of the disease sub-groups and these may be caused by alteration in metabolic precursors within the lung environment where addition by the phage offers an alternate pathway of metabolism. It is surprising to see the low level of association with "energy metabolism" in adult CF phage metagenomes and in >10 year clinical diagnosis nCFBR phage, but it may be possible that this is due to colonizing bacteria, in later stages of disease progression, may have evolved the perfect metabolism to survive and thus this relates to gene subtraction over addition. This also may be linked to the propensity of these Pa isolates to sustain in a biofilm.

The increase in the ability of certain phage communities to encode functionality that enables bacteria to degrade xenobiotic compounds is illustrated to evolve alongside the bacterium. It is notable that this also links to the previous reports of increasing antibiotic resistance in Pa isolated from the chronic lung (Winstanley et al., 2009). This pattern of increasing incidence of putative function also correlates to the increasing timeline of these diseases. This is illustrated by increases in incidence of metabolic pathways of cofactors and vitamins that have a role in generating biological activity. It may be possible that these cofactors are being encoded by the phage to aid bacterial survival and fitness. It must also be taken into account the apparent lack of these factors early on in these diseases may show a possible evolutionary timeline. The sizeable work undertaken here shows that adaptation of phage communities is apparent, with the amount of bacterium and possible number of phages offering confidence. This notable observation emphasizes the complexity and detailed nature of chronic lung infections and how phage evolution over time may affect the phenotype of Pa which in turn will have an overall impact on disease progression.

Bacteriophages evolve strategies alongside their bacterial host that promote infection, propagation and offer fitness to their host range. The inflamed lung is rich in mucus and the Pa isolated are variable in phenotype with some Pa expressing a mucoid surface. Therefore, it was pertinent to compare the BAM model of lytic phages as it may offer a way that temperate phages infect and transduce across their host range in the lower lung. Initially in this investigation we focused our attention on specific carbohydrate binding domains including Bacteriodetes Associated Carbohydrate Often N–terminal (BACON) (Mello et al., 2010) as this has been shown to be involved in the BAM model (Dutilh et al., 2014). However, searching Pa phage metagenomes using GeneWise2 identified no BACON domains in the panel of 92 phage samples. This does not refute the BAM model in this setting as the phage may bind to the Pa bacterial host via another carbohydrate or glycoprotein. Out of the 92 isolates and their associated phage, a Big_2 domain was observed in at least one of the phage's isolated from various clinical backgrounds (3 pediatric CF isolates, 15 adult CF isolates, 1 < 10 year clinical diagnosis nCFBR isolates and 9 > 10 year clinical diagnosis nCFBR isolates). When looking at the presence of the He_Pig motif it was identified that its occurrence was equal in both etiologies with a higher propensity in adult CF phage and >10 year Pa colonization nCFBR phage (17 and 14 respectively), putatively showing a role in phage adaptation. There were hits for the He_Pig domains in both pediatric CF phage (4) and <10 year clinical diagnosis nCFBR phage (7). Importantly we use three bioinformatics packages to assemble the DNA of each phage metagenome. We report that SPAdes and IBDA-UD are fairly comparable in performance, although we identified higher numbers of Big_2 domains using SPAdes. In addition Velvet optimizer yielded the poorest assembly when specifically targeting Ig-like domains with GeneWise2. Importantly though the Velvet derived assembly contained an Ig-domain (PF00047) that both SPAdes and IBDA-UD

failed to assemble. We would therefore recommend utilizing all 3 assemblers for future studies in this developing field.

Ig-like domains have been identified on approximately 25% of all sequenced *Caudovirales* genomes, there are three distinct families which are only identified in *Caudovirales* phage (Big_2, I-Set, and fn3) (Fraser et al., 2007). These domains on *Siphoviridae* and *Podoviridae* are located on Major Head, Major Tail and Tail Fiber proteins whilst in *Myoviridae* they are located on HOC, Fibritin and Baseplate proteins (Fraser et al., 2007). The lytic T4 phage in the BAM model shows the phage associating with mucus via the head HOC protein, this is seen with Electron Microscopy (EM) as the domain protrudes from the surface of the phage (Crusoe et al., 2010). Ig–like domains have been identified in the tail tube protein of *E. coli* bacteriophage λ, the exact reason for these Ig-like domains in λ is not fully known but when these domains are truncated, the phage has been seen to become more temperature sensitive (Katsura, 1981; Pell et al., 2010). These Ig–like domains may have accessory roles in bacterial infection rather than essential roles but their ubiquitous nature does potentially indicate there is an evolutionary advantage for the phage containing these domains (Pell et al., 2010; Barr et al., 2013). We show that Pa phages assembled in this study have a double Big_2 domain. The He_Pig domain was also seen in a double motif which flanked a putative minor phage tail protein (**Figure 6**). He_Pig is an Ig-like domain that is found in hemagglutinin and cell surface proteins, so it is possible that this domain is involved in the BAM model. The major sequence variations seen between the 4 types of the double Big_2 motif were found in the second Big_2 domain, **Figure 6B**. All these domains had a bit score above the proposed score of 25 and had a gap ranging from 18 to 6878 bp.

The BAM model reports on the role of the capsid-displayed, Ig-like protein domain of phage T4, and suggests a mechanism for its adherence to mucus. The putative carbohydrate-binding domains identified in this study appear to be associated with phage tail protein structures, rather than capsid domains. The presence of Ig-like domains therefore, on the phage isolates may indicate that the BAM model is functional for temperate phage but as a mode of infection and propagation rather than phage mediated immunity. We determine that rapid evolution may be making the overall diversity of phage Ig-like domains very large and so this may hinder the detection of these domains with GeneWise2 (Fraser et al., 2006) hence our suggestion of utilizing multiple assembly software. When comparing the clinical data associated with the bacterial isolates there is no correlation between the detection of Ig-like domains and the severity of disease.

This metagenomic project has indicated that the phage isolated from Pa derived from CF and nCFBR patients can associate with carbohydrates either on the bacterial host cells or in the mucus via Ig–like domains. When looking at the samples containing a Big_2 domain and/or a He_Pig domain, it was seen that they were more common in adult CF phage and >10 year clinical diagnosis nCFBR phage, possibly indicating that the presence of these domains has an evolutionary advantage for the phage's longevity in the chronic lung. This is the first time that this level of frequency has been observed for Ig–like domains in clinical isolates. There are multiple pathways which the Pa phage may utilize in order to adapt and offer fitness to its Pa host in the chronic lung environment (**Figure 7**). Through these data we can propose possible strategies the phages have adopted in order to sustain the ability to infect and propagate within

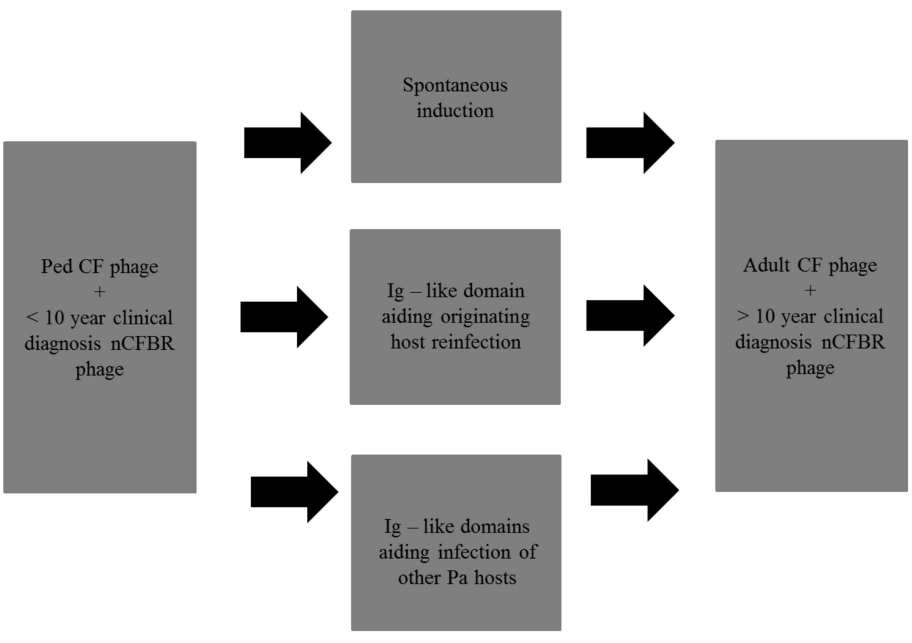


FIGURE 7 | Possible strategies that Pa phage can adapt over time in the chronic lung, various proposed mechanisms are included. Ped refers to pediatric CF patients in this figure and in future figures.

the lower lung, alongside etiological and evolutionary difference in the bacterial host. We have previously determined that temperate phages of Pa have high levels of spontaneous induction, with elevated levels being observed when Pa is in the early rather than mid-exponential growth phase. We hypothesize that phage are released from their host cell early to avoid becoming encased in a potential biofilm and have vegetative cells to infect. However, some of the phage that contains the Ig–like domains may aid adsorption to the cell from which they originate, or aid attachment and infection of other cells in their direct environment. Addition of traits such as adherence to the mucus lining the lower airways in order to increase their longevity in the chronic lung environment is also pertinent as an adaptive strategy.

This project has used metagenomics to show the array of adaptive mechanisms phage communities accrue over time in the lung environment. This work highlights the novel nature of metagenomics to understand complex communities without the need for a sensitive bacterial host or the necessity to enrich bacteriophage numbers and culture independence. We look at the "total arsenal" of a mixed phage community being induced from a clonal population and look at their possible impact on their next host in terms of infection and subversion. We map the effect that these mixed phage communities have on the functionality of their host or the strategies including BAM that have evolved over time. Using single phage lysates would not represent the complexity of the phage communities from each bacterium and their infection strategies in the chronic lung environment.

In conclusion, this study characterizes the inducible temperate phages found in Pa isolates sampled from the lungs of patients with CF and nCFBR. We provide the first evidence of how the complexities of the phage genomes possibly adapt over time and acquire accessory genes linking to key metabolic and signaling processes that may aid phage survival and bacterial fitness within the chronic lung. This study reports for the first time, to the best of our knowledge, the identification of multiple Ig-like domains on Pa temperate phages structural genes with such high frequency. There are increasing concerns over progressive anti-microbial resistance and the slow development pipeline for new antibiotics, so there is a real need for a developing novel paradigms and methods to overcome or further understand these challenges. These data offer temperate phages as a possible target in the resistance/ persistence pathways. Furthermore, they raise an additional clinical concern - could phage cross infection be a new challenge?

ACKNOWLEDGMENTS

We would like to acknowledge the sequencing facility NU-OMICS based at Northumbria University at Newcastle, UK). We are grateful to the National Institute for Health Research (NIHR) Local Clinical research network North East and North Cumbria for research nurse support. We would also like to thank all the nurses based both at the Freeman Hospital and Royal Victoria Infirmary, Newcastle Upon Tyne, UK for all their help collecting the Pa isolates involved in this investigation. We would also finally like to thank all of the patients that kindly agreed to participate in this scientific study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00097/abstract

REFERENCES

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., et al. (2013). Bacteriophage adhering to mucus provide a non–host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Beddoe, T., Paton, A. W., Le Nours, J., Rossjohn, J., and Paton, J. C. (2010). Structure, biological functions and applications of the AB5 toxins. *Trends Biochem. Sci.* 35, 411–418. doi: 10.1016/j.tibs.2010.02.003

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., and Nekrutenko, A. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26, 1783–1785. doi: 10.1093/bioinformatics/btq281

Boudewijns, M., Bakkers, J. M., Sturm, P. D., and Melchers, W. J. (2006). 16S rRNA gene sequencing and the routine clinical microbiology laboratory: a perfect marriage? *J. Clin. Microbiol.* 44, 3469–3470. doi: 10.1128/JCM.01017-06

Boyd, E. F., Carpenter, M. R., and Chowdhury, N. (2012). Mobile effector proteins on phage genomes. *Bacteriophage* 2, 139–148. doi: 10.4161/bact.21658

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv *P*reprint arXiv:1203.4802.

Crusoe, M. E., Edvenson, G., Fish, J., Howe, A., Irber, L., McDonald, E., et al. (2010). *khmer—k-mer Counting and Filtering FTW* [Online]. Available online at: http://khmer.readthedocs.org/en/latest/index.html [Accessed November 11, 2014].

Drancourt, M., Bollet, C., Carlioz, A., Martelin, R., Gayral, J.-P., and Raoult, D. (2000). 16S ribosomal DNA sequence analysis of a large collection of environmental and clinical unidentifiable bacterial isolates. *J. Clin. Microbiol.* 38, 3623–3630.

Dubreuil, J. D. (2012). The whole Shebang: the gastrointestinal tract, Escherichia coli enterotoxins and secretion. *Curr. Issues Mol. Biol.* 14, 71–82.

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5:4498. doi: 10.1038/ncomms5498

Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., and Salmond, G. P. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl. Acad. Sci. U.S.A.* 106, 894–899. doi: 10.1073/pnas.0808832106

Fraser, J. S., Maxwell, K. L., and Davidson, A. R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? *Curr. Opin. Microbiol.* 10, 382–387. doi: 10.1016/j.mib.2007.05.018

Fraser, J. S., Yu, Z., Maxwell, K. L., and Davidson, A. R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J. Mol. Biol.* 359, 496–507. doi: 10.1016/j.jmb.2006.03.043

Hansson, G. C. (2012). Role of mucus layers in gut infection and inflammation. *Curr. Opin. Microbiol.* 15, 57–62. doi: 10.1016/j.mib.2011.11.002

Hauser, A. R., Jain, M., Bar-Meir, M., and McColley, S. A. (2011). Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clin. Microbiol. Rev.* 24, 29–70. doi: 10.1128/CMR.00036-10

Hugenholtz, P., Goebel, B. M., and Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* 180, 4765–4774.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Katsura, I. (1981). Structure and function of the major tail protein of bacteriophage lambda: mutants having small major tail protein molecules in their virion. *J. Mol. Biol.* 146, 493–512. doi: 10.1016/0022-2836(81)90044-9

Kim, Y. S., and Ho, S. B. (2010). Intestinal goblet cells and mucins in health and disease: recent insights and progress. *Curr. Gastroenterol. Rep.* 12, 319–330. doi: 10.1007/s11894-010-0131-2

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., and Koonin, E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* 18, 11–19. doi: 10.1016/j.tim.2009.11.003

Lipuma, J. J. (2010). The changing microbial epidemiology in cystic fibrosis. *Clin. Microbiol. Rev.* 23, 299–323. doi: 10.1128/CMR.00068-09

Mall, M. A., and Boucher, R. C. (2014). Pathophysiology of cystic fibrosis lung disease. *Cystic Fibrosis* 64, 1. doi: 10.1183/1025448x.10008513

Matsushiro, A., Sato, K., Miyamoto, H., Yamamura, T., and Honda, T. (1999). Induction of prophages of enterohemorrhagic Escherichia coli O157:H7 with norfloxacin. *J. Bacteriol.* 181, 2257–2260.

McMahon, S. A., Miller, J. L., Lawton, J. A., Kerkow, D. E., Hodes, A., Marti-Renom, M. A., et al. (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat. Struct. Mol. Biol.* 12, 886–892. doi: 10.1038/nsmb992

Mello, L. V., Chen, X., and Rigden, D. J. (2010). Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Lett.* 584, 2421–2426. doi: 10.1016/j.febslet.2010.04.045

Muyzer, G., De Waal, E. C., and Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl. Environ. Microbiol.* 59, 695–700.

Nelson, A., De Soyza, A., Bourke, S. J., Perry, J. D., and Cummings, S. P. (2010). Assessment of sample handling practices on microbial activity in sputum samples from patients with cystic fibrosis. *Lett. Appl. Microbiol.* 51, 272–277. doi: 10.1111/j.1472-765X.2010.02891.x

Pell, L. G., Gasmi-Seabrook, G., Morais, M., Neudecker, P., Kanelis, V., Bona, D., et al. (2010). The solution structure of the C-terminal Ig-like domain of the bacteriophage λ tail tube protein. *J. Mol. Biol.* 403, 468–479. doi: 10.1016/j.jmb.2010.08.044

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15:76. doi: 10.1186/1471-2105-15-76

Rudkjobing, V. B., Thomsen, T. R., Alhede, M., Kragh, K. N., Nielsen, P. H., Johansen, U. R., et al. (2011). True microbiota involved in chronic lung infection of cystic fibrosis patients found by culturing and 16S rRNA gene analysis. *J. Clin. Microbiol.* 49, 4352–4355. doi: 10.1128/JCM.06092-11

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. New York, NY: Cold Spring Harbor Laboratory.

Smith, D. L., Rooks, D. J., Fogg, P. C., Darby, A. C., Thomson, N. R., McCarthy, A. J., et al. (2012). Comparative genomics of Shiga toxin encoding bacteriophages. *BMC Genomics* 13:311. doi: 10.1186/1471-2164-13-311

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., et al. (2005). Comparative metagenomics of microbial communities. *Science* 308, 554–557. doi: 10.1126/science.1107851

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Winstanley, C., Langille, M. G., Fothergill, J. L., Kukavica-Ibrulj, I., Paradis-Bleau, C., Sanschagrin, F., et al. (2009). Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool Epidemic Strain of *Pseudomonas aeruginosa*. *Genome Res.* 19, 12–23. doi: 10.1101/gr.086082.108

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

Received: 12 December 2014; accepted: 26 January 2015; published online: 18 February 2015.

*Citation: Tariq MA, Everest FLC, Cowley LA, De Soyza A, Holt GS, Bridge SH, Perry A, Perry JD, Bourke SJ, Cummings SP, Lanyon CV, Barr JJ and Smith DL (2015) A metagenomic approach to characterize temperate bacteriophage populations from Cystic Fibrosis and non-Cystic Fibrosis bronchiectasis patients. Front. Microbiol. **6**:97. doi: 10.3389/fmicb.2015.00097*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 January 2015
doi: 10.3389/fmicb.2015.00014

The human urine virome in association with urinary tract infections

Tasha M. Santiago-Rodriguez[1], Melissa Ly[1], Natasha Bonilla[2] and David T. Pride[1,3]*

[1] Department of Pathology, University of California, San Diego, San Diego, CA, USA
[2] Department of Biology, San Diego State University, San Diego, CA, USA
[3] Department of Medicine, University of California, San Diego, San Diego, CA, USA

Edited by:
Katrine L. Whiteson, University of California, Irvine, USA

Reviewed by:
Hiroyuki Sakai, Kyoto University, Japan
Lesley Ann Ogilvie, Alacris Theranostics GmbH/Max Planck Institute for Molecular Genetics, Germany

***Correspondence:**
David T. Pride, Department of Pathology, University of California, San Diego, 9500 Gilman Drive, MC 0612, La Jolla, San Diego, CA 92093-0612, USA
e-mail: dpride@ucsd.edu

While once believed to represent a sterile environment, the human urinary tract harbors a unique cellular microbiota. We sought to determine whether the human urinary tract also is home to viral communities whose membership might reflect urinary tract health status. We recruited and sampled urine from 20 subjects, 10 subjects with urinary tract infections (UTIs) and 10 without UTIs, and found viral communities in the urine of each subject group. Most of the identifiable viruses were bacteriophage, but eukaryotic viruses also were identified in all subjects. We found reads from human papillomaviruses (HPVs) in 95% of the subjects studied, but none were found to be high-risk genotypes that are associated with cervical and rectal cancers. We verified the presence of some HPV genotypes by quantitative PCR. Some of the HPV genotypes identified were homologous to relatively novel and uncharacterized viruses that previously have been detected on skin in association with cancerous lesions, while others may be associated with anal and genital warts. On a community level, there was no association between the membership or diversity of viral communities based on urinary tract health status. While more data are still needed, detection of HPVs as members of the human urinary virome using viral metagenomics represents a non-invasive technique that could augment current screening techniques to detect low-risk HPVs in the genitourinary tracts of humans.

Keywords: human microbiome, urinary tract infections, virome, virobiota, HPV, papillomavirus

INTRODUCTION

The presence of microbes inhabiting the human urinary tract has generally been associated with urinary tract infections (UTIs), but recent studies have demonstrated that the urine has its own unique microbiota even in the absence of UTIs (Nelson et al., 2010; Siddiqui et al., 2011; Wolfe et al., 2012). Many of these microbiota may not be culturable using conventional culture techniques, but the presence of a diverse microbiota in human urine has highlighted the need to understand whether there is a role for these communities of microbes in urinary tract health or disease. Healthy female urinary microbiota often include organisms also identified in the vagina (Fouts et al., 2012; Wolfe et al., 2012; Hilt et al., 2014), while healthy male microbiota may resemble the vagina, gut, and skin (Nelson et al., 2010; Dong et al., 2011; Fouts et al., 2012). There are numerous lower urinary tract symptoms that often do not have known infectious etiologies, so understanding whether these conditions may be influenced by disturbances to the urinary microbiota is of substantial interest.

UTIs are among the most common urological disorders in the health care settings, and are diagnosed using a combination of both symptoms and culture tests. *Escherichia coli* is recognized as the most common etiological agent of UTI, responsible for more than 80% of reported cases in females (Foxman and Brown, 2003), but there are numerous other microbes capable of causing UTIs (Stamm, 2002; Pour et al., 2011; Volkow-Fernandez et al., 2012). While a threshold concentration of 10^3 Colony Forming Units (CFU) per ml of urine accompanied with lower urinary tract symptoms is required for a UTI diagnosis, there are some pathogens that are not detected by standard microbiological culture techniques (Rubin et al., 1992). Disturbances to the urinary tract microbiota can be recognized in the presence of pathogens (Pearce et al., 2014), but it is not yet clear how long it takes for the urinary tract to recover its normal microbiota after both pathogen-mediated disturbances and antibiotic therapy.

While there is a substantial cellular microbiota indigenous to the human urinary tract, little is known about viruses. There are viral pathogens of the human urinary tract, such as adenovirus (Echavarria et al., 1998; Echavarria, 2008) and BK virus (Shinohara et al., 1993; Paduch, 2007; Egli et al., 2009), but these viruses generally are only pathogenic in subjects with compromised immune systems. There are other viruses such as Human papillomaviruses (HPVs) that inhabit human genital and rectal areas, and have previously also been found in urine (Pathak et al., 2014). Because HPVs can be associated with both warts and cancer in humans (Wylie et al., 2012), the diagnosis of its presence and particular subtypes associated with infection can be critical in the care of individual subjects who are infected. Previous studies have demonstrated that robust communities of cellular microbiota on human body surfaces are generally accompanied by communities of viruses. There are vast communities of viruses

that inhabit the human oral cavity (Pride et al., 2012; Abeles et al., 2014; Ly et al., 2014), the human gut (Minot et al., 2011), the human respiratory tract (Willner et al., 2009), and human skin (Foulongne et al., 2012), which suggests that there likely would be communities of viruses that also are indigenous to the human urinary tract. One prior study identified a unique viral community in the subgingival crevice of individuals with severe periodontal disease (Ly et al., 2014), which suggests that viruses could play a role in health and disease. There are many subjects with lower urinary tract symptoms, with no known cellular microbial etiologies, yet viral communities in these subjects are completely unexplored (Enerly et al., 2013).

We sought to decipher whether there is a population of viruses indigenous to the human urinary tract and whether urine viral community membership may be affected by urinary tract health status. Our goals were to: (1) identify and quantify viral populations in human urine, (2) determine whether trends in urine bacterial biota might also be reflected in the virobiota of urine, (3) elucidate whether the presence of urinary pathogens affects viral community membership, (4) decipher whether human viruses populate the urinary tract, or whether viruses encountered are primarily bacteriophage, and (5) characterize differences between the virobiota of the male and female urinary tracts.

MATERIALS AND METHODS

HUMAN SUBJECTS AND CULTURE CONDITIONS

Human subject involvement in this study was approved by the University of California, San Diego Administrative Panel on Human Subjects in Medical Research. The study was certified as category 4 exempt, which includes research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if the information is recorded in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects. We sampled the urine from 20 human subjects, with 10 having urinary tract infections, and another 10 testing negative for UTIs. None of the female subjects had prior abnormal PAP smears and therefore had never been tested for the presence of cervical HPV. None of the male subjects had anal PAP smears and were not previously tested for HPV. No subjects reported any prior history of anal or genital warts. The diagnosis of UTI was based on the clinical laboratory standards institute definition, and includes the presence of $\geq 10^3$ CFU of bacteria and $\geq$10 leukocytes per high powered field (Stamm, 1983). All specimens were planted on Sheep's blood agar, McConkey, and Chocolate agar plates. Each of the patients in this study was symptomatic and had at least 10^3 CFU/ml of bacteria in their urine specimens. Bacteria were identified to the species level using Matrix-Assisted Laser Desorption Ionization Time of Flight mass spectrometry, which is a mass spectrometry technique for the identification of bacteria with sensitivity and specificity similar to that of conventional biochemical techniques (Neville et al., 2011). For the *E. coli* isolates, their identification also was verified by the presence of lactose fermentation and mobility to ensure that they could not be *Shigella* species. Urine specimens were processed within 2 h of their collection and the bacteria were identified within 24 h of their collection. All specimens from both UTI+ and UTI− subjects had their viromes processed within 48 h of their collection. All cultures from UTI− subjects were held for a minimum of 72 h to ensure they were truly negative. All urine specimens were stored at 4°C prior to processing of viromes.

VIROME PREPARATION AND SEQUENCING

Urine samples were processed according to our previously described protocols for processing viruses from saliva (Pride et al., 2012). Urine (1 ml) was filtered sequentially using 0.45 and 0.2 μm filters (VWR, Radnor, PA) and purified on a cesium chloride density gradient, with the fraction corresponding to most known bacteriophage (Murphy et al., 1995) retained. The purified virions were further purified on Amicon YM-100 protein columns (Millipore, Inc., Bellerica, MA), treated with DNase I, and their DNA purified using a Qiagen UltraSens virus kit (Qiagen, Valencia, CA). DNA then was amplified using GenomiPhi V2 MDA amplification (GE Healthcare, Pittsburgh, PA), fragmented to 200–400 bp using a Bioruptor (Diagenode, Denville, NJ), and libraries created using the Ion Plus Fragment Library Kit. Libraries then were sequenced using 314 chips on an Ion Torrent Personal Genome Machine (PGM; Life Technologies, Grand Island, NY) (Rothberg et al., 2011) producing an average read length of approximately 206 bp.

VIROME ANALYSIS

We trimmed the resulting reads according to modified Phred scores of 0.5 using CLC Genomics Workbench 4.65 (CLC bio USA, Cambridge, MA), removed any low complexity reads (where >25% of the length were due to homopolymer tracts), and removed any reads with substantial length variation (<50 nucleotides or >300 nucleotides) or ambiguous characters prior to further analysis. Each virome was screened for contaminating bacterial and human nucleic acids using BLASTN analysis (E-score $<10^{-5}$) against the Ribosomal Database Project 16S rRNA database (Cole et al., 2009), and the human reference database available at ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/. Any reads homologous to human sequences were removed prior to further analysis. GC content variation among contigs was assessed using Box and Whiskers plots created using Microsoft Excel 2007 (Microsoft Corp., Redman, WA). Remaining reads were assembled using CLC Genomics Workbench 4.65 based on 98% identity with a minimum of 50% read overlap, which were more stringent than criteria developed to discriminate between highly related viruses (Breitbart et al., 2002). Because the shortest reads were 50 nucleotides, the minimum tolerable overlap was 25 nucleotides, and the average overlap was no less than 103 nucleotides depending on the characteristics of each virome. The consensus sequence for each contig was constructed according to majority rule, and any contigs <200 nucleotides or with ambiguous characters were removed prior to further analysis. Contigs were annotated using BLASTX against the NCBI Non-redundant (NR) database with an E-score cutoff value of 10^{-5}. Specific viral homologs were determined by parsing BLASTX results for known viral genes including replication, structural, transposition, restriction/modification, hypothetical, and other genes previously found in viruses for which the E-score was at least 10^{-5}. Each individual virome contig was annotated using this technique; however, if the best hit for any portion of the contig was to a gene with no known

function, lower level hits were used as long as they had known function and still met the E-score cutoff. The annotation data were compiled for each subject and used to determine the relative proportions of assembled contigs that contained viral homologs. Analysis of shared homologs present in each virome was performed by creating custom BLAST databases for each virome, comparing each database with all other viromes using BLASTN analysis (E-score $<10^{-10}$). Beta diversity between viromes was determined using binary Sorensen distances, and was determined by randomly sampling 1000 contigs between viromes. Principal coordinates analysis (PCOA) was performed with the estimated distances using Qiime (Caporaso et al., 2010b). Read mappings of viromes to a database of phage (www.phantome.org; ftp://ftp.ncbi.nih.gov/genomes/Viruses/) and a database of viruses (ftp://ftp.ncbi.nih.gov/genomes/Viruses/) was performed using CLC Genomics Workbench 4.65 (CLC bio USA, Cambridge, MA), and were mapped using 98% identity over a minimum of 50% of the read length. Virome sequences are available for download in the MG-RAST database (metagenomics.anl.gov/) under the project "UrineViromeProject," or under project #9680.

VIRAL DIVERSITY

To measure alpha diversity in the viral communities, we utilized a technique termed the Homologous Virus Diversity Index (HVDI). The technique is based on finding high levels of homology amongst contigs within viromes that likely belong to the same virus but were placed into separate contigs due to the limitations of the assembly process (Abeles et al., 2014). Virome reads were assembled using 98% identify over a minimum of 50% of the read length using CLC Genomics Workbench 4.65 (CLC bio USA, Cambridge, MA), and the resulting contig spectra utilized as the primary input for the index. We created custom nucleotide BLAST databases for each subject that contained all their contigs. We then used BLASTN analysis to find high levels of homology (E-score $<10^{-20}$) between different contigs within the same subject. We accepted only high levels of homology that spanned at least 50% of the length of the shorter contig being compared. All contigs in each subject were treated as nodes and those contigs that had high homology to other contigs in the same subject were added to a network by directing edges between the nodes. After evaluating homologies among all intra-subject contigs, networks formed from directed edges/nodes were assigned to individual viruses and nodes with no associations were considered singular viruses. For each resulting network, we added the number of reads assigned to each node on the network and the combined number of reads was used to represent the relative abundance of the virus represented by that network. The relative abundances of all viruses were calculated using this technique, and a new contig spectra representing the viral population in each subject was formed. The contig spectra from each subject then were used as surrogates for population structures and input directly into the Shannon Index (Gotelli and Colwell, 2001) to estimate diversity. We also utilized the relative numbers of viral contigs created from only a single read (singleton) or from only two reads (doubleton) as input for the chao1 index (Chao, 1984). The results of the corrected contig spectra in each case were compared with uncorrected contig spectra, and in each case resulted in an approximate 10% reduction in diversity calculations.

QUANTITATIVE PCR

We designed primers HPV49F (GTGATTCAAATGCAAGGG) and HPV49R (TTTACAATCTCAGACCAGTG) to target HPV genotype 49 and HPV178F (TGTTTGGAAGGGATCTAGT) and HPV178R (TCATTAAAAAGCGCCAGG) to target HPV genotype 178. Primers for both genotypes were based on early gene 1 sequences. We performed quantitative PCR in 20 μl reactions using 10 μl of SYBR Select Master Mix (Life Technologies, Grand Island, NY), 1 μl of template viral DNA (10 ng/μl), 7 μl of ultrapure water, and 1 μl (0.5 pmol) of each of the forward and reverse primers. The following cycling parameters were used: an initial denaturation at 95°C for 2 min, followed by 35 cycles at 95°C for 15 s, annealing at 54°C for 15 s, and final extension at 60°C for 1 min on an Eppendorf Mastercycler Realplex2. Each reaction was run in triplicate and results were only considered positive if each replicate had a *Ct*-value =35 with a product at the expected size. One of the products was verified by Sanger sequencing of the PCR amplicon and produced a product identical to the reads matching HPV already found.

ANALYSIS OF 16S rRNA

Genomic DNA was prepared from the urine of each subject and time point using the Qiagen QIAamp DNA MINI kit (Qiagen, Valencia, CA). Each sample was subjected to a bead beating step prior to nucleic acid extraction using Lysing Matrix-B (MP Bio, Santa Ana, CA). We amplified the bacterial 16S rRNA V3 hypervariable region using the forward primer 341F (CCTACGGGAGGCAGCAG) fused with the Ion Torrent Adaptor A sequence and one of 23 unique 10 base pair barcodes, and reverse primer 514R (ATTACCGCGGCTGCTGG) fused with the Ion Torrent Adaptor P1 from the urine of each subject (Whiteley et al., 2012). PCR reactions were performed using Platinum PCR SuperMix (Invitrogen, Carlsbad, CA) with the following cycling parameters: 94°C for 10 min, followed by 30 cycles of 94°C for 30 s, 53°C for 30 s, 72°C for 30 s, and a final elongation step of 72°C for 10 min. Resulting amplicons were purified on a 2% agarose gel stained with SYBR Safe (Invitrogen, Carlsbad, CA) using the MinElute PCR Purification kit (Qiagen, Valencia, CA). Amplicons were further purified with Ampure beads (Beckman-Coulter, Brea, CA), and molar equivalents were determined for each sample using a Bioanalyzer 2100 HS DNA Kit (Agilent Technologies, Santa Clara, CA). Samples were pooled into equimolar proportions and sequenced on 314 chips using an Ion Torrent PGM according to manufacturer's instructions (Life Technologies, Grand Island, NY) (Rothberg et al., 2011). Resulting sequence reads were removed from the analysis if they were <130 nucleotide, had any barcode or primer errors, contained any ambiguous characters, or contained any stretch of >6 homopolymers. Sequences were assigned to their respective samples based on their 10 nucleotide barcode sequence, and were analyzed further using the Qiime pipeline (Caporaso et al., 2010b). Briefly, representative OTUs from each set were chosen at a minimum sequence identity of 97% using UClust (Edgar, 2010) and aligned using PyNast (Caporaso et al., 2010a) against

the Greengenes database (Desantis et al., 2006). Multiple alignments then were used to create phylogenies using FastTree (Li and Godzik, 2006), and taxonomy was assigned to each OTU using the RDP classifier (Wang et al., 2007; Price et al., 2009). PCOA was performed based on Beta Diversity using weighted Unifrac distances (Lozupone et al., 2006). Alpha diversity based on the Shannon Index and the chao1 index also were performed using the Qiime pipeline. Differences in the relative abundances of taxa between subject groups were determined using student's *t*-test. 16S rRNA sequences are available for download in the MG-RAST database (metagenomics.anl.gov/) under the project "UrineViromeProject," or under project #9680.

STATISTICAL ANALYSIS

To assess whether viromes had significant overlap between subjects and subject groups, we performed a permutation test based on resampling (10,000 iteration) (Robles-Sikisaka et al., 2013, 2014; Abeles et al., 2014; Ly et al., 2014; Naidu et al., 2014). We simulated the distribution of the fraction of shared virome homologs from two different time points within individual subjects that were randomly chosen across all time points. For each set, we computed the summed fraction of shared homologs using 1000 random contigs between randomly chosen individual time points within different subjects, and from these computed an empirical null distribution of our statistic of interest (the fraction of shared homologs). The simulated statistics within each subject group were referred to the null distribution of inter-group comparisons, and the *p*-value was computed as the fraction of times the simulated statistic for the each exceeded the observed statistic. An identical analysis was performed at the OTU level for the16S rRNA taxonomic assignments. For analysis of sex specific characteristics within the viromes, a randomly chosen subject and time point from the male sex was compared with a randomly chosen subject and time from the female sex to determine the null distribution of fraction of shared contigs based on opposite sexes. We then estimated the fraction of shared homologs from randomly chosen subjects and time points within each sex and compared with the empirical null distribution from simulated inter-sex values. We estimated the *p*-value based on the fraction of times the intra-sex statistic exceeded that for the observed statistic.

RESULTS

HUMAN SUBJECT CHARACTERISTICS

We recruited 20 human subjects, 10 that were diagnosed with UTIs and 10 with negative urine cultures (**Table 1**). The diagnosis of UTI was made in the urine based upon the presence of $\geq$10 leukocytes per high powered field along with $\geq 10^3$ CFU of bacteria. Of those 10 subjects with UTI's, eight grew *E. coli*, one grew *Enterococcus faecium*, and one grew *Acinetobacter baumanii* complex from their urine. Five of the subjects were male and five were female in both arms of the study. We processed viromes from the urine in each subject within 24 h of their collection to ensure that viral structure and nucleic acid integrity would not be affected by processing times.

EPIFLUORESCENCE MICROSCOPY

Because viral communities in human urine have not previously been characterized, we first utilized epifluorescence microscopy to determine whether viral communities were present and to estimate their concentrations. We found that there were approximately 10^7 virus-like particles (VLPs) per ml of urine present in each of the subjects in this study (Supplemental Figure 1). No differences in VLP concentrations were observed between subjects with or without UTIs. Comparatively, there generally are 10^8 VLPs per ml of saliva (Pride et al., 2012) and 5×10^8 VLPs per ml of human feces (Haynes and Rohwer, 2011). The relatively low concentrations of viruses in urine may reflect the lower relative abundance of the microbiota in the human urinary tract when compared to the oral cavity and the gut. These data suggest that there are communities of viruses present in the human urinary tract whose relative abundance are not affected by the presence of urinary pathogens.

URINE VIROME SEQUENCE CHARACTERISTICS

We isolated viral communities from the urine of our cohort using our previously described methods (Robles-Sikisaka et al., 2013, 2014; Abeles et al., 2014; Ly et al., 2014; Naidu et al., 2014). We sequenced a total of 14,542,349 reads, 6,734,864 from subjects with UTIs and 7,807,485 from subjects without UTIs (Supplemental Table 1). The mean number of reads per subject was 727,117, with a mean length of 206 nucleotides, and a mean GC content of 40.1%. Viromes were screened for contaminating cellular DNA by BLASTN (E-score $<10^{-5}$) analyses against the Ribosomal Database Project 16S rRNA database (Cole et al., 2009) and a Human Reference Genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/). None of the viromes had identifiable 16S rRNA, and 0.42% of the reads were homologous to human DNA. Reads homologous to human DNA were removed prior to further analysis. We assembled the viral reads into contigs, as the larger contigs generally result in more productive BLAST searches. We obtained an average of 2285 contigs per subject with a mean length of 1209 nucleotides and a mean GC content of 41.4%. No significant differences were observed in GC content between subjects with and without UTIs (Supplemental Figure 2).

DETECTION OF HUMAN PAPILLOMAVIRUSES (HPVs)

To decipher whether there may be individual viruses in this cohort that have high similarity to known viruses, we mapped the reads from each virome to the NCBI virus database. We found reads in both UTI+ and UTI− subjects that matched herpesviruses, polyomaviruses, and human papillomaviruses. The HPVs were much more prevalent than any other eukaryotic viruses identified in these subjects, and many of the viromes had reads that mapped specifically across most of the genomes of the HPVs. Nineteen of the 20 (95%) subjects mapped across different HPV viromes, which included all UTI− subjects and nine of 10 (90%) of the UTI+ subjects (Supplemental Table 2). For subject URN2, 1.7% (14,647 reads) of the reads mapped to HPV Type 96 (**Figure 1A**), and in subject URN6, 1.8% (13,948 reads) of the virome reads mapped to HPV Type 49 (**Figure 1B**). Different reads from these subjects also mapped to HPV Types 92, and 121 (data not shown). In subject URP12, 3.7% (25,389 reads) of the reads mapped to HPV Type 178 (**Figure 1C**) and Type 121 (data not shown). None of the subjects participating in this study had previously been diagnosed with genital warts or HPV infections, so the presence

Table 1 | Study subjects.

Subjects	Sex	Age	Symptoms	Immuno compromised[a]	Organism	Catheter[b]
NEGATIVE URINE CULTURES						
URN1	Male	94	None	No	None	No
URN2	Male	57	Sepsis[c]	Yes	None	Yes
URN6	Male	51	Pain in prostate area	No	None	Yes
URN9	Male	47	Pain with urination	No	None	No
URN10	Female	27	Vaginal bleeding	No	None	No
URN11	Female	31	Pain with urination	No	None	No
URN12	Female	25	Abdominal pain	No	None	No
URN13	Female	51	Urine odor	Yes	None	No
URN15	Female	54	Sepsis[c]	No	None	Yes
URN16	Male	72	None	No	None	No
POSITIVE URINE CULTURES						
URP1	Male	62	Urinary retention	No	*E. coli*	No
URP2	Female	34	Sepsis[c]	Yes	*E. coli*	Yes
URP3	Female	50	Stroke	No	*E. coli*	No
URP4	Male	65	None	No	*E. coli*	Yes
URP6	Female	18	Dysuria	No	*E. coli*	No
URP7	Female	22	Autonomic instability	No	*E. faecium*	Yes
URP9	Male	52	Sepsis[c]	Yes	*E. coli*	No
URP10	Male	58	Sepsis[c]	Yes	*Acinetobacter*	Yes
URP12	Male	78	Sepsis[c]	Yes	*E. coli*	Yes
URP14	Female	69	Dysuria	Yes	*E. coli*	No

[a] *Includes subjects with cancer and organ transplants, and those taking immunosuppressive medications such as steroids.*

[b] *Includes subjects with catheters in their bladders and a single subject with a nephrostomy.*

[c] *Potentially life threatening complication of infection resulting in severe inflammation and blood pressure drops.*

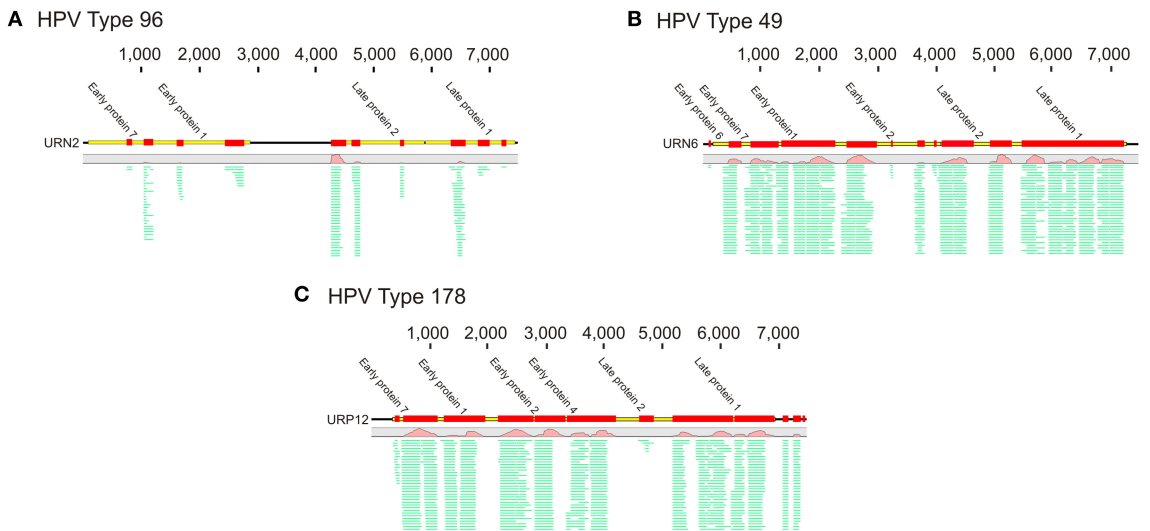


FIGURE 1 | Read mappings of select viromes to Human papillomaviruses (HPVs). (A) Represents URN2 mapping to HPV Type 96, **(B)** represents URN6 mapping to HPV Type 49, and **(C)** represents URP12 mapping to HPV Type 178. The relative coverage of each contig is represented, along with annotated open reading frames (ORFs) above each box. The portions of the contigs identified are represented by the colored boxes for each subject. The length of each contig is denoted at the top of each panel. The relative proportions of reads mapping to each virus are represented in green.

of HPV in the urine of these subjects suggests that urine may be a reasonable specimen type for screening for urogenital HPV infections. None of the females had previously had abnormal PAP smears or cervical HPV testing, but the types of HPVs identified by commercially available tests would not have identified the genotypes present in this study. To confirm the presence of some HPV types in the urine, we developed primers for HPV Types 49 and 178, as several subjects had reads that mapped specifically

to these viruses. We confirmed the presence of nucleic acids matching HPV Type 49 in the urine of five of the 20 subjects (Supplemental Figure 3), while 10% of the subjects had nucleic acids matching HPV Type 178 (**Figure 2**).

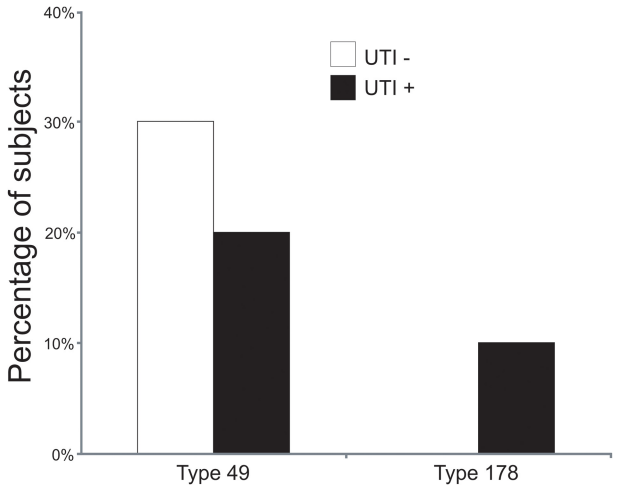


FIGURE 2 | Bar graphs of the percentage of subjects with detectable HPV Types 49 and 178 by quantitative PCR. White bars indicate those subjects with negative urine cultures, and black bars represent subjects with urinary tract infections.

IDENTIFICATION OF URINARY BACTERIOPHAGE

Viromes were also mapped against a phage database. We found many viral reads that mapped to individual viruses, however, many only mapped to single genes often involved in viral replication. We did find some that mapped across much of the genomes of known viruses, including subject URN2 mapping to Lambda phage (Supplemental Figure 4A), URN12 mapping to *Staphylococcus* phage PH15 (Supplemental Figure 4B), URP1 mapping to *E. coli* phage phiV10 (Supplemental Figure 4C), and URP2 mapping to *Enterococcus* phage phiFL4A (Supplemental Figure 4D).

To further aid in deciphering the contents of the viromes, we subjected all the assembled contigs to BLASTX analysis against the NCBI NR database to identify homologous sequences. We identified numerous contigs with homologous sequences in the NR database, with most of those sequences homologous to phage. Approximately 27 percent of the contigs were homologous to known viruses, with the majority (>99%) representing bacteriophage (Supplemental Figure 5). We identified contigs with homologs to viral hypothetical proteins, structural genes (head, capsid, collar, portal tail, tail fibers), restriction modification enzymes, virulence and DNA replication and integration genes (DNA polymerases, helicases, integrases, and primases) (**Figure 3**). No significant differences were noted for the viral gene categories in association with urological health status. The high proportions of integrases identified (**Figure 3**) suggests that viruses with primarily lysogenic lifestyles were abundant in the urine viromes.

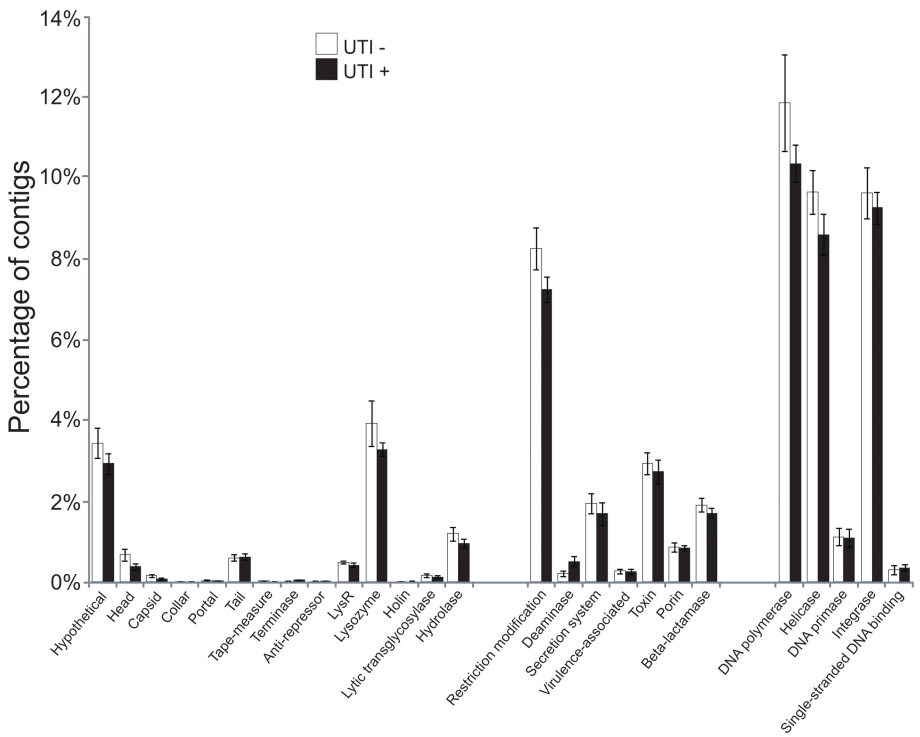


FIGURE 3 | Bar graphs of the mean percentages of contigs (± standard error) with viral homologs in the NR database from all of the subjects. White bars indicate those subjects with negative urine cultures, and black bars represent subjects with urinary tract infections.

ALPHA DIVERSITY AMONGST URINARY MICROBIOTA

We determined the alpha diversity present in the urinary viromes to determine if significant differences exist between subjects based on urinary tract health status and as a measure of the relative adequacy of our sequencing depth. We performed rarefaction analysis on the urinary viromes based on the Homologous Virus Diversity Index (HVDI) (Santiago-Rodriguez et al., under review) to determine the adequacy of our sequencing depth. We first calculated the HVDI based on the Shannon Index (**Figure 4A**), which indicated that most of the diverse viruses present in the community had been sampled with relatively minimal sequencing efforts. The index calculated based on the chao1 index (**Figure 4B**) approached asymptote for most urine samples, which corroborates that the sequencing depth for these urinary viromes was adequate.

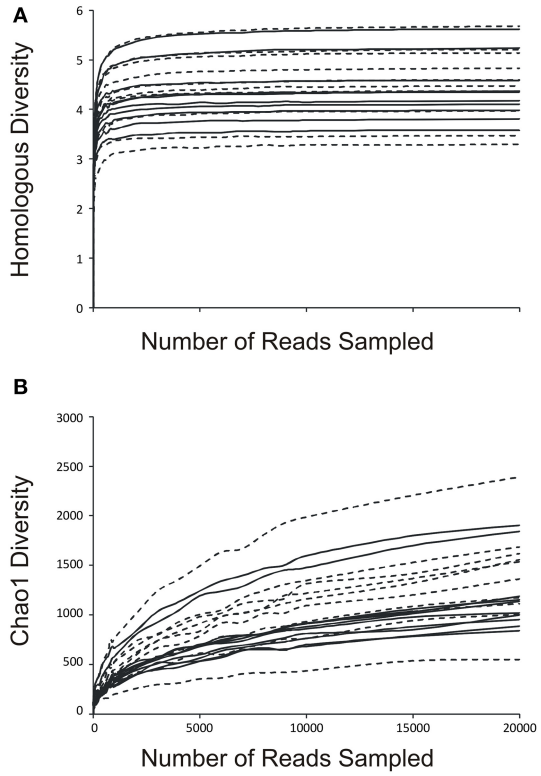


FIGURE 4 | Rarefaction analysis based on the Homologous Virus Diversity Index using the Shannon Index (A) and the chao1 Index (B) for urine viral communities. The index values are shown on the y-axis and the number or reads sampled are shown on the x-axis. White boxes indicate those subjects with negative urine cultures, and black triangles represent subjects with urinary tract infections. Dashed lines indicate those subjects with negative urine cultures, and solid lines represent subjects with urinary tract infections.

We also characterized the bacterial communities through analysis of the 16S rRNA V3-hypervariable region. We compared the alpha diversity present in bacterial and viral communities using the Shannon index for the bacterial biota. There were no observed differences in viral community diversity based on urinary tract health status (**Figure 5A**) ($p = 0.729$); however, bacterial diversity was substantially lower in subjects with UTIs (**Figure 5B**) ($p = 0.059$), probably reflecting the relative abundance of the urinary pathogens present. That viral diversity is unaffected despite differences in bacterial biota, suggests that the urinary pathogens may not contribute substantially to overall viral diversity.

BETA DIVERSITY IN URINARY MICROBIOMES

To decipher whether there may exist differences in virome community membership based on urinary tract health status, we compared beta diversity amongst the viromes and visualized patterns of variation present using principal coordinates analysis (**Figure 6A**). We did not observe any patterns of variation in the viromes that were attributable to health status. We did, however, identify substantial differences when examining the bacterial biota, where those subjects with UTIs formed a somewhat homogenous cluster (**Figure 6B**). We previously have identified host-sex specific characteristics of viral communities in human saliva (Abeles et al., 2014), and we tested whether there may be specific traits of the virobiota of urine that might also be host-sex specific. For the viral communities, no host-sex specific differences were identified for viral communities (Supplemental Figure 6).

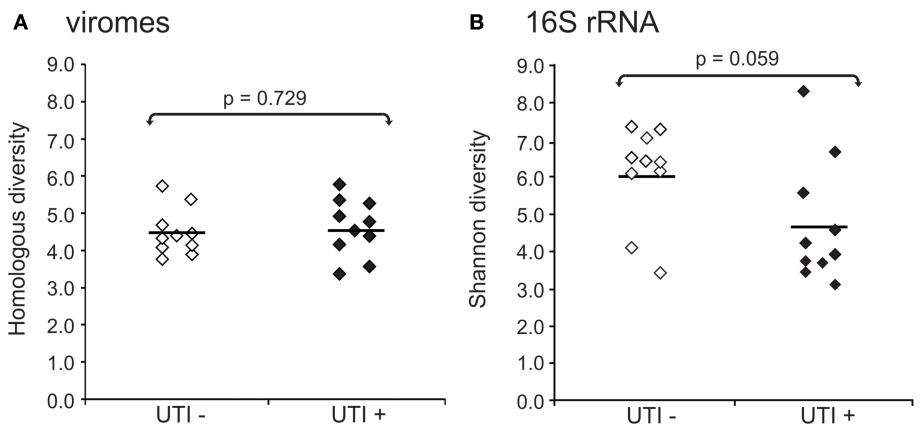


FIGURE 5 | Plots of alpha diversity for urine viromes (A) and bacteria biota using 16S rRNA (B). Viral diversity was determined using the Homologous Virus Diversity Index based on the Shannon Index, and bacterial diversity determined using the Shannon Index. White diamonds indicate those subjects with negative urine cultures, and black diamonds represent subjects with urinary tract infections.

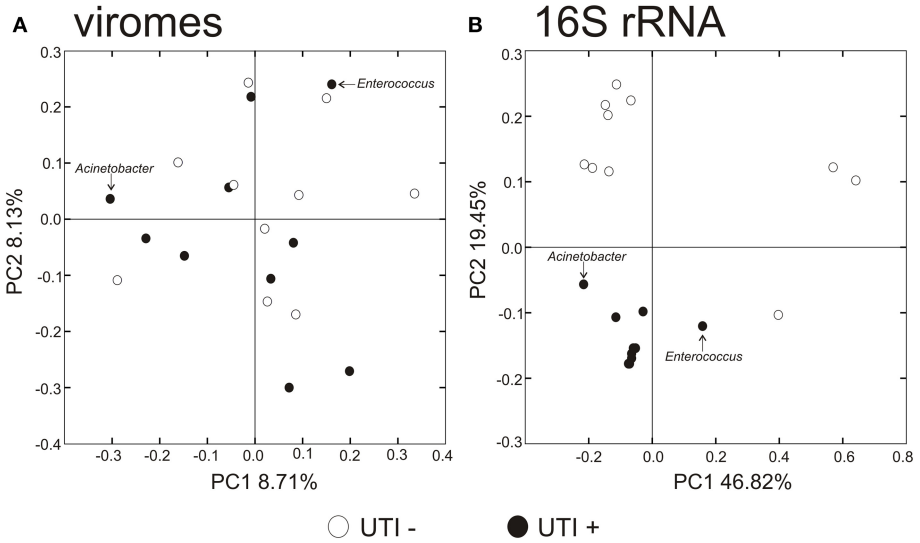


FIGURE 6 | Principal coordinates analysis of beta diversity of the urine viromes (A), and bacterial biota using 16S rRNA (B). White circles indicate those subjects with negative urine cultures, and black circles represent subjects with urinary tract infections.

Table 2 | Viral homologs and shared 16S OTUs within and between subject groups.

	Percent homologous within group[a]	Percent homologous between groups[a]	*p*-Value[b]
BY URINE HEALTH STATUS			
Viromes			
UTI−	24.14 ± 4.86	23.73 ± 4.34	0.453
UTI+	23.79 ± 4.51	23.70 ± 4.32	0.512
16S rRNA			
UTI−	84.02 ± 19.05	73.36 ± 27.23	0.343
UTI+	87.29 ± 16.36	73.60 ± 27.12	0.301
BY SEX			
Viromes			
Male	23.28 ± 3.93	23.64 ± 4.44	0.524
Female	24.66 ± 4.63	23.74 ± 4.48	0.424
16S rRNA			
Male	88.47 ± 10.82	75.42 ± 25.99	0.377
Female	76.55 ± 27.72	76.26 ± 25.48	0.470

[a] *Based on the mean of 10,000 iterations. Thousand random contigs were sampled per iteration.*

[b] *Empirical p-value based on the fraction of times the estimated percent homologous contigs or shared OTUs for each group exceeded that between groups.*

We utilized a permutation test to determine whether the proportion of homologous viral contigs among subjects with UTIs was greater than would be expected to occur by chance (Abeles et al., 2014; Ly et al., 2014). Similar to the results for the PCOA (**Figure 6A**), there were no significant differences observed for virome community membership by urinary tract health status (**Table 2**). While the differences were not statistically significant, there were more shared bacterial OTUs amongst the subjects based on their urinary tract health status. There were no significant differences for viromes or the bacterial biota based on host sex.

BLAST HITS AND BACTERIAL TAXONOMIES

We characterized the BLASTX hits for each virus homolog by the taxonomy of the host bacteria at the Phylum level. While this type of characterization does not provide accurate identification of the putative hosts for each viral homolog, it can be utilized comparatively to understand whether virome homologs might reflect 16S rRNA taxonomies. By characterizing the viral communities in this manner, we find that BLASTX hits to Proteobacteria are most predominant, followed by Bacteroidetes, Actinobacteria, Verrucomicrobia, and Firmicutes (**Figures 7A,B** and Supplemental Figure 7). The BLASTX hits were similar regardless of urinary tract health status, which again suggests that urinary tract health does not determine virome community membership. Trends in urine viral communities were not represented in the bacterial taxonomies (**Figures 7C,D**).

DISCUSSION

Human body surfaces are inhabited by indigenous microbial communities whose role in health and disease are just beginning to be elucidated. The urine was once believed to constitute a sterile environment, but studies now have determined that both the male and female genitourinary tracts are home to viable bacterial communities. We hypothesized that because these communities are home to cellular microbes that they would also have indigenous communities of viruses, and this study is the first to demonstrate that urine indeed has a robust community of viruses. We found that there were approximately 10^7 VLPs in the urine of UTI+ and UTI− subjects, which is an order of magnitude lower than we have previously found in human saliva (Pride et al., 2012). This may reflect differences in the relative abundances of bacteria in the urine compared to the oral cavity, as the healthy human oral cavity has many cultivable bacterial biota, while few are found in healthy

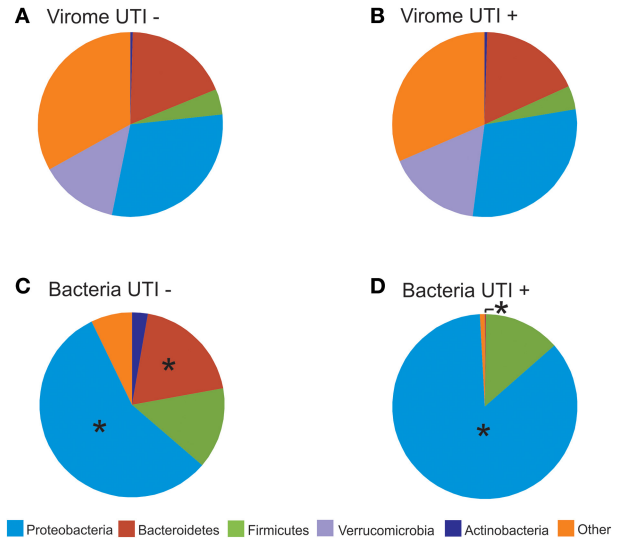


FIGURE 7 | Pie charts of phage BLASTX hits (A,B) and bacterial taxonomy using 16S rRNA (C,D). (A) and **(C)** represent subjects with negative urine cultures, and **(B)** and **(D)** represent subjects with urinary tract infections. The "∗" represents significant differences ($p = 0.05$) between the proportions of different phyla represented between UTI+ and UTI− subjects.

urine. While many of the viruses identified were bacteriophage, as has been demonstrated in many prior studies (Breitbart et al., 2003; Willner et al., 2009, 2011; Reyes et al., 2010; Minot et al., 2012; Pride et al., 2012; Ly et al., 2014; Naidu et al., 2014; Robles-Sikisaka et al., 2014), the presence of HPVs, which are eukaryotic viruses of the human host, suggests that human urine is a complex community with numerous different indigenous virus genotypes. Our methods to characterize viral communities generally have not identified many enveloped viruses; however, we previously have identified herpesviruses in the human oral cavity using these methods (Pride et al., 2012; Ly et al., 2014), so the extent of any biases imposed by using CsCl density gradient centrifugation remains unclear. We focused our analysis of phage on the contigs assembled from the virome reads because the longer contigs allowed for more productive searches for homologous sequences, but 73% of the contigs had no identifiable homologs. The analysis of contigs did not identify many eukaryotic viruses because their relatively low abundance and probable low diversity resulted in few assembled contigs.

Many of our studies have found few eukaryotic viruses in the oral microbiome when compared to the abundance and diversity of bacteriophage (Pride et al., 2012; Robles-Sikisaka et al., 2013; Abeles et al., 2014; Ly et al., 2014), suggesting that eukaryotic viruses represent a relatively small proportion of the human virome (Wylie et al., 2012). In the present study we showed that HPVs comprise a portion of the urine virome in some subjects regardless of urinary tract health status. We were able to detect HPV Types in 19 of the 20 (95%) subjects studied, suggesting that HPV may be relatively common urinary virobiota. The HPV genotypes identified in this study have not been associated with high risks for cervical carcinomas, but Type 9 identified in some subjects may be associated with the development of cutaneous cancers (Andersson et al., 2012). Other subjects mapped to Type 49, which has been isolated from warts in renal transplant patients (Favre et al., 1989), and Type 178, which was isolated from skin adjacent to actinic keratosis (Johansson and Forslund, 2014; Martin et al., 2014). Our results are among the first to demonstrate the presence of these HPV genotypes in urine, presenting the opportunity to further investigate their role in urological health and diverse urological diseases. Unfortunately, the cohort of subjects studied was not ideal for characterizing potential pathogenic effects of the HPV genotypes identified because none had prior cystoscopies, culposcopies, or body imaging to document bladder or cervical pathology. Because the HPV genotypes in this study were not high risk genotypes, cystoscopies and culposcopies were not clinically indicated. Typical protocols for identifying HPV in women are based on the presence of atypical cervical pathology (Nobbenhuis et al., 1999), although primary HPV screening is now becoming more prevalent (Whitlock et al., 2011). None of the females in this study had prior abnormal PAP smears, and thus had not been tested for HPV infections. HPV diagnostic tests in clinical laboratories test for high risk genotypes (Ronco et al., 2010), and would not have identified the genotypes found in this study. We believe that most of the genotypes we identified probably may be associated with genital warts, although none of the subjects previously reported having genital warts.

We characterized the human urinary virome in association with UTIs, as they are the most frequent cause of genitourinary morbidity in adults (Resnick and Older, 1997; Foxman, 2003). It was somewhat surprising that the bacterial communities were significantly altered in UTI+ subjects (**Figures 6**, **7**), while there were no obvious differences observed in urine viral communities. In the human oral cavity, we previously found that both bacterial and viral communities were altered in subjects with moderate/severe periodontal disease (Ly et al., 2014). However, even when viral communities are composed mostly of bacteriophage, their membership does not necessarily reflect that of their host communities, likely due to different dynamic relationships present for different host/phage pairs (Pride et al., 2012). We believe that the lack of differences observed in UTI+ subjects may be secondary to the pathogens in these infections contributing relatively few viruses to the urinary microbiota. Further work would be necessary to decipher the contribution of phage from the pathogens in these subjects' urine microbiomes.

Similar to previous studies, there were identifiable differences in bacteria taxa in association with urological health status (Pearce et al., 2014). Urine from UTI− subjects exhibited higher bacterial diversity when compared to UTI+ subjects (**Figure 5B**). These differences were not observed for viral communities (**Figure 5A**), as the pattern of BLASTX homologs in UTI+ and UTI− subjects suggests that much of the viral community is conserved regardless of urinary tract health status (**Figures 7A,B**). Recent work has demonstrated that phage are capable of attaching to mucosal surfaces (Barr et al., 2013), so it is possible that the viruses observed in UTI+ patients could reflect viruses swept into the urine from mucosal surfaces rather than actively replicating urine

viruses. Study of the urinary tract transcriptome in a cohort of UTI+ and UTI− patients could help elucidate whether the viruses observed were actively transcribing/replicating.

The majority of the studies characterizing the urine microbiome in association with urological disorders focus on female subjects, as lower urinary tract symptoms are more common in adult females than males. Given anatomical differences in male and female urinary tracts, it is feasible to hypothesize that there would also be differences in the urinary microbiota. While not meeting statistical significance, it was interesting that there were differences identified in the bacterial biota of males and females regardless of infection status (**Table 2** and Supplemental Figure 6). Much of the differences may reflect the uniqueness of the female genitourinary tract, where the vagina also has its own unique microbiota (Srinivasan et al., 2010) that potentially overlaps with the urinary tract. Unlike our prior studies in the human oral cavity (Abeles et al., 2014), there was no association between host-sex and urine viral community membership (**Table 2** and Supplemental Figure 6). This study included only 20 subjects, which we used to screen for potential differences in virome contents between the sexes. It remains a possibility that a much larger cohort could identify differences urine viromes differences based on host sex. There appeared to be substantial inter-individual variability between the viromes of the subjects studied, however, we believe that longitudinal rather than cross-sectional studies of human viromes provide a better overview of variation observed between individuals (Abeles et al., 2014).

Despite specific HPV types being associated with low and high-health risks, just a fraction of the over 170 types have been directly implicated with disease, and many may not be detected using routine methods (Chaturvedi et al., 2011). Negative results may lead to undiagnosed illnesses such as cancerous lesions; yet, negative results have also been implicated with the existence of novel HPV types that remain to be characterized and associated with diverse disease phenotypes (Johansson et al., 2013). High-throughput sequencing has been shown to successfully detect several different HPV types in HPV-negative condylomas, suggesting that it may be a reasonable approach to detect HPV in HPV-negative specimens (Johansson et al., 2013). Our detection of HPV genotypes in the urine viromes of 19 subjects whom previously had not been diagnosed with HPV infections suggests that HPV colonization of the urine may be relatively common. That most viromes mapped across different HPVs, suggests the viruses in our subjects share common features with multiple known HPV genotypes. The data presented in this study indicate that viral metagenomics on urine could augment current HPV diagnostic approaches (Pathak et al., 2014).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00014/abstract

REFERENCES

Abeles, S. R., Robles-Sikisaka, R., Ly, M., Lum, A. G., Salzman, J., Boehm, T. K., et al. (2014). Human oral viruses are personal, persistent and gender-consistent. *ISME J.* 8, 1753–1767. doi: 10.1038/ismej.2014.31

Andersson, K., Michael, K. M., Luostarinen, T., Waterboer, T., Gislefoss, R., Hakulinen, T., et al. (2012). Prospective study of human papillomavirus seropositivity and risk of nonmelanoma skin cancer. *Am. J. Epidemiol.* 175, 685–695. doi: 10.1093/aje/kwr373

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., et al. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185, 6220–6223. doi: 10.1128/JB.185.20.6220-6223.2003

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Caporaso, J. G., Bittinger, K., Bushman, F. D., Desantis, T. Z., Andersen, G. L., and Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267. doi: 10.1093/bioinformatics/btp636

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Chao, A. (1984). Non-parametric estimation of the number of classes in a population. *Scand. J. Statist.* 11, 265–270.

Chaturvedi, A. K., Katki, H. A., Hildesheim, A., Rodríguez, A. C., Quint, W., Schiffman, M., et al. (2011). Human papillomavirus infection with multiple types: pattern of coinfection and risk of cervical disease. *J. Infect. Dis.* 203, 910–920. doi: 10.1093/infdis/jiq139

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145. doi: 10.1093/nar/gkn879

Desantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072. doi: 10.1128/AEM.03006-05

Dong, Q., Nelson, D. E., Toh, E., Diao, L., Gao, X., Fortenberry, J. D., et al. (2011). The microbial communities in male first catch urine are highly similar to those in paired urethral swab specimens. *PLoS ONE* 6:e19709. doi: 10.1371/journal.pone.0019709

Echavarria, M. (2008). Adenoviruses in immunocompromised hosts. *Clin. Microbiol. Rev.* 21, 704–715. doi: 10.1128/CMR.00052-07

Echavarria, M., Forman, M., Ticehurst, J., Dumler, J. S., and Charache, P. (1998). PCR method for detection of adenovirus in urine of healthy and human immunodeficiency virus-infected individuals. *J. Clin. Microbiol.* 36, 3323–3326.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Egli, A., Infanti, L., Dumoulin, A., Buser, A., Samaridis, J., Stebler, C., et al. (2009). Prevalence of polyomavirus BK and JC infection and replication in 400 healthy blood donors. *J. Infect. Dis.* 199, 837–846. doi: 10.1086/597126

Enerly, E., Olofsson, C., and Nygard, M. (2013). Monitoring human papillomavirus prevalence in urine samples: a review. *Clin. Epidemiol.* 5, 67–79. doi: 10.2147/CLEP.S39799

Favre, M., Obalek, S., Jablonska, S., and Orth, G. (1989). Human papillomavirus type 49, a type isolated from flat warts of renal transplant patients. *J. Virol.* 63, 4909.

Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M. A., et al. (2012). Human skin microbiota: high diversity of DNA viruses identified on the human skin by high throughput sequencing. *PLoS ONE* 7:e38499. doi: 10.1371/journal.pone.0038499

Fouts, D. E., Pieper, R., Szpakowski, S., Pohl, H., Knoblach, S., Suh, M. J., et al. (2012). Integrated next-generation sequencing of 16S rDNA and metaproteomics differentiate the healthy urine microbiome from asymptomatic bacteriuria in neuropathic bladder associated with spinal cord injury. *J. Transl. Med.* 10:174. doi: 10.1186/1479-5876-10-174

Foxman, B. (2003). Epidemiology of urinary tract infections: incidence, morbidity, and economic costs. *Dis. Mon.* 49, 53–70. doi: 10.1067/mda.2003.7

Foxman, B., and Brown, P. (2003). Epidemiology of urinary tract infections: transmission and risk factors, incidence, and costs. *Infect. Dis. Clin. North Am.* 17, 227–241. doi: 10.1016/S0891-5520(03)00005-9

Gotelli, N. J., and Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4, 379–391. doi: 10.1046/j.1461-0248.2001.00230.x

Haynes, M., and Rohwer, F. (2011). "The human virome," in *Metagenomics of the Human Body* (New York, NY: Springer), 63–77. doi: 10.1007/978-1-4419-7089-3_4

Hilt, E. E., McKinley, K., Pearce, M. M., Rosenfeld, A. B., Zilliox, M. J., Mueller, E. R., et al. (2014). Urine is not sterile: use of enhanced urine culture techniques to detect resident bacterial flora in the adult female bladder. *J. Clin. Microbiol.* 52, 871–876. doi: 10.1128/JCM.02876-13

Johansson, H., Bzhalava, D., Ekstrom, J., Hultin, E., Dillner, J., and Forslund, O. (2013). Metagenomic sequencing of "HPV-negative" condylomas detects novel putative HPV types. *Virology* 440, 1–7. doi: 10.1016/j.virol.2013.01.023

Johansson, H., and Forslund, O. (2014). Complete genome sequences of three novel human papillomavirus types, 175, 178, and 180. *Genome Announc.* 2:e00443-14. doi: 10.1128/genomeA.00443-14

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

Lozupone, C., Hamady, M., and Knight, R. (2006). UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 7:371. doi: 10.1186/1471-2105-7-371

Ly, M., Abeles, S. R., Boehm, T. K., Robles-Sikisaka, R., Naidu, M., Santiago-Rodriguez, T., et al. (2014). Altered oral viral ecology in association with periodontal disease. *MBio* 5:e01133-14. doi: 10.1128/mBio.01133-14

Martin, E., Dang, J., Bzhalava, D., Stern, J., Edelstein, Z. R., Koutsky, L. A., et al. (2014). Characterization of three novel human papillomavirus types isolated from oral rinse samples of healthy individuals. *J. Clin. Virol.* 59, 30–37. doi: 10.1016/j.jcv.2013.10.028

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3962–3966. doi: 10.1073/pnas.1119061109

Minot, S., Sinha, R., Chen, J., Li, H., Keilbaugh, S. A., Wu, G. D., et al. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 21, 1616–1625. doi: 10.1101/gr.122705.111

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., Martelli, G. P., et al. (1995). *Virus Taxonomy: Sizth Report of the International Committee on Taxonomy of Viruses*, Vol. Suppl. 10. New York, NY: Springer-Verlag.

Naidu, M., Robles-Sikisaka, R., Abeles, S. R., Boehm, T. K., and Pride, D. T. (2014). Characterization of bacteriophage communities and CRISPR profiles from dental plaque. *BMC Microbiol.* 14:175. doi: 10.1186/1471-2180-14-175

Nelson, D. E., Van Der Pol, B., Dong, Q., Revanna, K. V., Fan, B., Easwaran, S., et al. (2010). Characteristic male urine microbiomes associate with asymptomatic sexually transmitted infection. *PLoS ONE* 5:e14116. doi: 10.1371/journal.pone.0014116

Neville, S. A., Lecordier, A., Ziochos, H., Chater, M. J., Gosbell, I. B., Maley, M. W., et al. (2011). Utility of matrix-assisted laser desorption ionization-time of flight mass spectrometry following introduction for routine laboratory bacterial identification. *J. Clin. Microbiol.* 49, 2980–2984. doi: 10.1128/JCM.00431-11

Nobbenhuis, M. A., Walboomers, J. M., Helmerhorst, T. J., Rozendaal, L., Remmink, A. J., Risse, E. K., et al. (1999). Relation of human papillomavirus status to cervical lesions and consequences for cervical-cancer screening: a prospective study. *Lancet* 354, 20–25. doi: 10.1016/S0140-6736(98)12490-X

Paduch, D. A. (2007). Viral lower urinary tract infections. *Curr. Urol. Rep.* 8, 324–335. doi: 10.1007/s11934-007-0080-y

Pathak, N., Dodds, J., Zamora, J., and Khan, K. (2014). Accuracy of urinary human papillomavirus testing for presence of cervical HPV: systematic review and meta-analysis. *BMJ* 349:g5264. doi: 10.1136/bmj.g5264

Pearce, M. M., Hilt, E. E., Rosenfeld, A. B., Zilliox, M. J., Thomas-White, K., Fok, C., et al. (2014). The female urinary microbiome: a comparison of women with and without urgency urinary incontinence. *MBio* 5:e01283-14. doi: 10.1128/mBio.01283-14

Pour, N. K., Dusane, D. H., Dhakephalkar, P. K., Zamin, F. R., Zinjarde, S. S., and Chopade, B. A. (2011). Biofilm formation by *Acinetobacter baumannii* strains isolated from urinary tract infection and urinary catheters. *FEMS Immunol. Med. Microbiol.* 62, 328–338. doi: 10.1111/j.1574-695X.2011.00818.x

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Pride, D. T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R. A. III, et al. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J.* 6, 915–926. doi: 10.1038/ismej.2011.169

Resnick, M. I., and Older, R. A. (1997). *Diagnosis of Genitourinary Disease.* New York, NY: Thieme.

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., et al. (2010). Viruses in the faecal microbiota of monozytotic twins and their mothers. *Nature* 466, 334–338. doi: 10.1038/nature09199

Robles-Sikisaka, R., Ly, M., Boehm, T., Naidu, M., Salzman, J., and Pride, D. T. (2013). Association between living environment and human oral viral ecology. *ISME J.* 7, 1710–1724. doi: 10.1038/ismej.2013.63

Robles-Sikisaka, R., Naidu, M., Ly, M., Salzman, J., Abeles, S. R., Boehm, T. K., et al. (2014). Conservation of streptococcal CRISPRs on human skin and saliva. *BMC Microbiol.* 14:146. doi: 10.1186/1471-2180-14-146

Ronco, G., Giorgi-Rossi, P., Carozzi, F., Confortini, M., Dalla Palma, P., Del Mistro, A., et al. (2010). Efficacy of human papillomavirus testing for the detection of invasive cervical cancers and cervical intraepithelial neoplasia: a randomised controlled trial. *Lancet Oncol.* 11, 249–257. doi: 10.1016/S1470-2045(09)70360-2

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352. doi: 10.1038/nature10242

Rubin, R. H., Shapiro, E. D., Andriole, V. T., Davis, R. J., and Stamm, W. E. (1992). Evaluation of new anti-infective drugs for the treatment of urinary tract infection. Infectious Diseases Society of America and the Food and Drug Administration. *Clin. Infect. Dis.* 15(Suppl. 1), S216–S227. doi: 10.1093/clind/15.Supplement_1.S216

Shinohara, T., Matsuda, M., Cheng, S. H., Marshall, J., Fujita, M., and Nagashima, K. (1993). BK virus infection of the human urinary tract. *J. Med. Virol.* 41, 301–305. doi: 10.1002/jmv.1890410408

Siddiqui, H., Nederbragt, A. J., Lagesen, K., Jeansson, S. L., and Jakobsen, K. S. (2011). Assessing diversity of the female urine microbiota by high throughput sequencing of 16S rDNA amplicons. *BMC Microbiol.* 11:244. doi: 10.1186/1471-2180-11-244

Srinivasan, S., Liu, C., Mitchell, C. M., Fiedler, T. L., Thomas, K. K., Agnew, K. J., et al. (2010). Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. *PLoS ONE* 5:e10197. doi: 10.1371/journal.pone.0010197

Stamm, W. E. (1983). Measurement of pyuria and its relation to bacteriuria. *Am. J. Med.* 75, 53–58. doi: 10.1016/0002-9343(83)90073-6

Stamm, W. E. (2002). Scientific and clinical challenges in the management of urinary tract infections. *Am. J. Med.* 113(Suppl. 1A), 1S–4S. doi: 10.1016/S0002-9343(02)01053-7

Volkow-Fernandez, P., Rodriguez, C. F., and Cornejo-Juarez, P. (2012). Intravesical colistin irrigation to treat multidrug-resistant *Acinetobacter baumannii* urinary tract infection: a case report. *J. Med. Case Rep.* 6:426. doi: 10.1186/1752-1947-6-426

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Whiteley, A. S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., et al. (2012). Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *J. Microbiol. Methods* 91, 80–88. doi: 10.1016/j.mimet.2012.07.008

Whitlock, E. P., Vesco, K. K., Eder, M., Lin, J. S., Senger, C. A., and Burda, B. U. (2011). Liquid-based cytology and human papillomavirus testing to screen for cervical cancer: a systematic review for the U.S. Preventive Services Task Force. *Ann. Intern. Med.* 155, 687–697, W214–W685. doi: 10.7326/0003-4819-155-10-201111150-00376

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., et al. (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4:e7370. doi: 10.1371/journal.pone.0007370

Willner, D., Furlan, M., Schmieder, R., Grasis, J. A., Pride, D. T., Relman, D. A., et al. (2011). Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc. Natl. Acad. Sci. U.S.A.* 108(Suppl. 1), 4547–4553. doi: 10.1073/pnas.1000089107

Wolfe, A. J., Toh, E., Shibata, N., Rong, R., Kenton, K., Fitzgerald, M., et al. (2012). Evidence of uncultivated bacteria in the adult female bladder. *J. Clin. Microbiol.* 50, 1376–1383. doi: 10.1128/JCM.05852-11

Wylie, K. M., Weinstock, G. M., and Storch, G. A. (2012). Emerging view of the human virome. *Transl. Res.* 160, 283–290. doi: 10.1016/j.trsl.2012.03.006

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

Received: 20 November 2014; paper pending published: 03 December 2014; accepted: 06 January 2015; published online: 23 January 2015.

Citation: Santiago-Rodriguez TM, Ly M, Bonilla N and Pride DT (2015) The human urine virome in association with urinary tract infections. Front. Microbiol. ***6****:14. doi: 10.3389/fmicb.2015.00014*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 March 2015
doi: 10.3389/fmicb.2015.00216

Identification of staphylococcal phage with reduced transcription in human blood through transcriptome sequencing

Tasha M. Santiago-Rodriguez[1], *Mayuri Naidu*[1], *Marcus B. Jones*[2], *Melissa Ly*[1] and *David T. Pride*[1,3]*

[1] Department of Pathology, University of California, San Diego, CA, USA, [2] J. Craig Venter Institute, Rockville, MD, USA, [3] Department of Medicine, University of California, San Diego, CA, USA

Many pathogenic bacteria have bacteriophage and other mobile genetic elements whose activity during human infections has not been evaluated. We investigated the gene expression patterns in human subjects with invasive Methicillin Resistant *Staphylococcus aureus* (MRSA) infections to determine the gene expression of bacteriophage and other mobile genetic elements. We developed an *ex vivo* technique that involved direct inoculation of blood from subjects with invasive bloodstream infections into culture media to reduce any potential laboratory adaptation. We compared *ex vivo* to *in vitro* profiles from 10 human subjects to determine MRSA gene expression in blood. Using RNA sequencing, we found that there were distinct and significant differences between *ex vivo* and *in vitro* MRSA gene expression profiles. Among the major differences between *ex vivo* and *in vitro* gene expression were virulence/disease/defense and mobile elements. While transposons were expressed at higher levels *ex vivo*, lysogenic bacteriophage had significantly higher *in vitro* expression. Five subjects had MRSA with bacteriophage that were inhibited by the presence of blood in the media, supporting that the lysogeny state was preferred in human blood. Some of the phage produced also had reduced infectivity, further supporting that phage were inhibited by blood. By comparing the gene expression cultured in media with and without the blood of patients, we gain insights into the specific adaptations made by MRSA and its bacteriophage to life in the human bloodstream.

Keywords: *Staphylococcus aureus*, transcriptome, RNA Seq, human microbiome, bacteriophage, prophage, mobile genetic element

OPEN ACCESS

Edited by:
Katrine L. Whiteson, University of California, USA

Reviewed by:
Theo Dreher, Oregon State University, USA
Beatriz Martínez, Consejo Superior de Investigaciones Científicas, Spain
Jodi A. Lindsay, St George's, University of London, UK

***Correspondence:**
David T. Pride, Department of Pathology, University of California, San Diego, 9500 Gilman Drive, MC 0612, La Jolla, San Diego, CA 92093-0612, USA
dpride@ucsd.edu

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *21 November 2014*
Accepted: *03 March 2015*
Published: *24 March 2015*

Citation:
Santiago-Rodriguez TM, Naidu M, Jones MB, Ly M and Pride DT (2015) Identification of staphylococcal phage with reduced transcription in human blood through transcriptome sequencing. Front. Microbiol. 6:216. doi: 10.3389/fmicb.2015.00216

Introduction

Staphylococcus aureus is a pathogen that is also considered normal human flora, and often takes advantage of breaks in protective skin barriers to cause disease (Chaffin et al., 2012). While *S. aureus* strains were primarily treatable with beta-lactam antibiotics in the past, their widespread use has resulted in the emergence of Methicillin-Resistant *S. aureus* (MRSA) strains. MRSA can be acquired in hospital- or community-based settings, but many of the Community-Acquired MRSA (CA-MRSA) strains have replaced more traditional Hospital-Acquired (HA-MRSA) strains in both environments (Shopsin et al., 2003; Popovich et al., 2008; Kennedy et al., 2010). CA-MRSA generally has been responsible for many invasive soft tissue infections, and has several virulence

factors including Panton-Valentine leukocidin (PVL), which are thought to contribute greatly to its pathogenesis. PVL, which is not regularly detected in Hospital Acquired (HA)-MRSA strains, have been associated with epidemics of several CA-MRSA strains in the United States (Pan et al., 2003; Vandenesch et al., 2003). Virulence factors are of particular interest in invasive bloodstream infections, and many are derived from mobile genetic elements, including plasmids, bacteriophage, transposons, and pathogenicity islands (Bae et al., 2006; Baba et al., 2008; Diep et al., 2008b).

There usually are bacteriophage integrated into the genomes of *S. aureus* isolates, with most of them belong to the phage family Siphoviridae (Canchaya et al., 2003; Feng et al., 2008). These phage are of intermediate size (generally around 40 to 45 kb) and often carry toxins that may contribute to pathogenesis (Brussow et al., 2004). Some phage carry the immune modulator staphylokinase, which is responsible for host tissue destruction. Others encode toxins such as PVL, or superantigens involved in toxic shock syndrome, necrotizing fasciitis, and food poisoning (Deghorain and Van Melderen, 2012). These phage likely impact staphylococcal pathogenesis through lysogenic conversions, where the virulence functions they carry are expressed during infection in humans. The expression of these phage may directly reflect their contributions to pathogenesis, but has not been thoroughly examined during human bloodstream infections.

The expression of genes involved in the pathogenicity of CA-MRSA has been mainly studied *in vitro* (Cui et al., 2005; Lindsay et al., 2006; Stevens et al., 2007; Pohl et al., 2009). Few studies have compared the *in vivo* and *in vitro* gene expression of MRSA, and generally have been restricted to animal models (Diep et al., 2008a,b; Chaffin et al., 2012), which may not fully reflect the adaptive behavior of the pathogen in humans. Other *ex vivo* strategies have grown a single, lab-adapted MRSA strain in the presence of healthy human donor blood to characterize virulence genes that may be overexpressed (Malachowa and Deleo, 2011; Malachowa et al., 2011). No studies, however, have studied MRSA across human subjects with varying degrees of illnesses to understand whether the behavior of MRSA and their mobile genetic elements is reproducible across different human subjects. It is of substantial importance to understand the contributions of phage and other mobile genetic elements to MRSA pathogenesis during human infections and their potential for spread to others.

The primary tool for analysis of global patterns of gene expression in microbes is transcriptomics. While some have studied transcriptomics in MRSA utilizing quantitative PCR (Sabersheikh and Saunders, 2004), or microarray analysis (Witney et al., 2005; Lindsay et al., 2006), RNA sequencing provides a tool for characterization of patterns of gene expression that does not require a priori information about the pathogen being studied (Wilhelm and Landry, 2009). Thus, gene expression patterns from RNA sequencing can be used to thoroughly characterize novel MRSA strains. RNA sequencing technology has not been employed to characterize global changes in *S. aureus* gene expression in humans with invasive bloodstream infections. Here, we report the RNA sequencing expression profiles of MRSA strains grown in the presence of blood from 10 human subjects with invasive bloodstream infections and characterize differences observed in the expression of bacteriophage and other mobile genetic elements by comparing expression profiles with and without human blood.

Results

Human Subjects and RNA Enrichment

We sampled blood from 10 human subjects with invasive bloodstream MRSA infections (**Table 1** and Supplemental Table 1), and cultured each to further our understanding of the gene expression of MRSA in the human bloodstream. We utilized an *ex vivo* technique that involved direct inoculation at the bedside of blood from human subjects with invasive bloodstream infections into culture media that bypassed the need for a separate *in vitro* culture step; thus, the isolates from each subject were characterized directly from blood with minimal opportunity for gene expression changes that might accompany laboratory adaptation. The cohort of subjects in this study included many that were critically-ill (**Table 1**), which significantly contrasts with a prior study characterizing a single lab-adapted MRSA strain cultured in the presence of blood from healthy donors (Malachowa et al., 2011). For comparison, we performed a separate isolation of each MRSA strain, and cultured each *in vitro* to help decipher comparatively those genes whose expression might be induced through exposure to the human bloodstream. Samples 21MRA and 23MRA were isolated from the same individual 48 h apart, and were utilized to help determine whether results would be consistent within individual subjects over time. All *ex vivo* and *in vitro* MRSA isolates were grown to log phase (**Supplemental Figure 1**), and total RNA was isolated from each subject/MRSA isolate under both growth conditions. Because the *ex vivo* samples were cultured in the presence of blood from each human subject, we enriched to remove any RNA that may have been derived from the human host. The resulting RNA was sequenced from all 10 subjects for a total of 6,213,492 *ex vivo* reads (mean of 564,863 $\pm$ 79,540 per subject) and 4,568,420 *in vitro* reads (mean of 415,310 $\pm$ 10,582 per subject) (**Table 2**). The inclusion of the enrichment step substantially reduced the presence of human RNA, as exemplified by the relatively low percentage of RNA identified that was homologous to human DNA (mean 1.22 $\pm$ 0.55%; range from 0.17% to 6.26%). All sequence reads homologous to human DNA were removed prior to further analysis.

Identification of Staphylococcal Sequence Reads

We mapped the sequence reads from each subject to a database of known staphylococcal genomes (available at ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) to determine whether we could identify gene sequences from the *Staphylococcus* genome that were expressed upon exposure to human blood. We found that the vast majority of the *ex vivo* sequenced reads mapped to known staphylococcal genomes (mean 98.78 $\pm$ 0.32%), with a similar proportion of the *in vitro* sequenced reads (98.06 $\pm$ 0.19%) also mapping to known staphylococcal genomes. That a similar percentage of *ex vivo* and *in vitro* sequenced reads mapped to staphylococcal genomes, indicates

TABLE 1 | Study subjects.

	Age	Sex	Diagnosis	Comorbidities	Antibiotics[a]
1MRA	64	Male	Bacteremia; severe sepsis; abscess	Quadriplegia; hypertension	Vancomycin/Gentamicin/Rifampin
3MRA	57	Female	Bacteremia; sepsis	Metastatic colon cancer	Vancomycin
4MRA	84	Female	Bacteremia; sepsis; infected dialysis graft	End-stage renal disease	Vancomycin/Gentamicin/Televancin
6MRA	64	Female	Bacteremia; septic arthritis	Metastatic colon cancer	Vancomycin
20MRA	60	Male	Bacteremia; sepsis; furunculosis	Asthma	Vancomycin
21/23MRA[b]	50	Male	Bacteremia; sepsis; line infection	Dermatomyositis	Vancomycin/Daptomycin
31MRA	60	Female	Bacteremia; severe sepsis	Schleroderma	Vancomycin/Piperacillin/Tazobactam
55MRA	60	Female	Bacteremia; sepsis; line infection	Congestive heart failure	Vancomycin/Rifampin
74MRA	81	Male	Bacteremia; sepsis; Pneumonia	High blood pressure	Vancomycin/Piperacillin/Tazobactam
256MRA	73	Female	Bacteremia; sepsis; osteomyelitis	T12 paraplegia; sacral decubitus ulcer	Vancomycin/Piperacillin/Tazobactam

[a] *Antibiotics administered within 24 h prior to the sample collection.*
[b] *Subjects 21 and 23 represent the same individual for which samples were collected at different time points.*

TABLE 2 | RNA sequences from all subjects.

	Reads	Average length	Homologous to human genome	After trimming	Reads mapping to *S.aureus* (%)	Top genome
EX VIVO						
1 MRA	505,449	103	27,413	437,829	427,847 (97.72)	*S. aureus* COL
3 MRA	1,266,504	106	3447	1,166,335	1,164,602 (99.85)	*S. aureus* COL
4 MRA	1,023,554	105	490	880,026	878,868 (99.87)	*S. aureus* JH1
6 MRA	630,865	104	11,848	575,675	571,640 (99.30)	*S. aureus* USA300
20 MRA	336,473	105	1432	289,966	287,842 (99.27)	*S. aureus* JH1
*21 MRA	766,395	107	8330	681,855	674,333 (98.90)	*S. aureus* USA300
*23 MRA	583,685	104	9419	464,855	459,591 (98.87)	*S. aureus* USA300
31 MRA	657,978	100	3419	596,716	595,028 (99.72)	*S. aureus* COL
55 MRA	489,670	96	585	298,828	289,039 (96.72)	*S. aureus* JH1
74 MRA	478,942	107	1129	410,085	405,586 (98.90)	*S. aureus* USA300
256 MRA	443,073	108	71	411,322	400,666 (97.41)	*S. aureus* USA300
IN VITRO						
1 MRA	479,683	105	2	404,995	397,930 (98.26)	
3 MRA	532,094	110	1	481,800	466,891 (96.91)	
4 MRA	474,698	108	0	414,247	407,187 (98.30)	
6 MRA	454,227	111	1	414,825	402,384 (97.00)	
20 MRA	442,793	105	2	403,797	399,087 (98.83)	
*21 MRA	440,061	106	8	408,518	401,531 (98.29)	
*23 MRA	466,089	109	2	444,640	434,200 (97.65)	
31 MRA	465,569	107	1	392,500	386,082 (98.36)	
55 MRA	401,893	100	4	357,369	349,760 (97.87)	
74 MRA	514,879	110	4	460,010	453,521 (98.59)	
256 MRA	449,617	107	2	385,719	380,309 (98.60)	

[*] *Subjects 21 and 23 represent the same individual for which samples were collected at different time points.*

the enrichment process was highly specific for bacterial RNA. Four of the 10 subjects likely harbored MRSA USA300 type strains based on their high percentage of mapping reads, while the other six subjects harbored strains that mapped to more traditional HA-MRSA strain types. Both samples 21MRA and 23MRA mapped to USA300 strains, indicating that this strain type was identical over the 48 h between samplings in that individual subject.

Ex Vivo vs. *In Vitro* MRSA Gene Expression Profiles

We examined the overall gene expression from each subject both *ex vivo* and *in vitro* to determine whether there were global differences in gene expression that might signal adaptations to human blood. We found that there was a significant difference in the proportion of MRSA genes expressed, with far fewer genes expressed *ex vivo* (47.88 ± 5.39% *ex vivo* vs. 60.82 ±

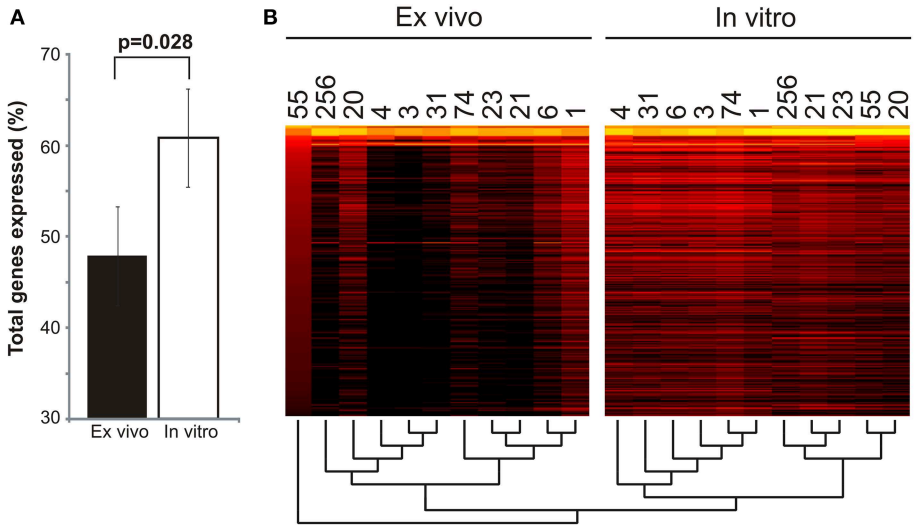


FIGURE 1 | Global gene expression of *Staphylococcus aureus* cultured in blood from human subjects (*ex vivo*) and *in vitro*. **(A)**—mean percentage (± standard error) of total genes expressed *ex vivo* and *in vitro*. **(B)** —heatmap of gene expression profiles *ex vivo* and *in vitro* for all subjects. Each column represents an individual subject, and each row represent each individual *S. aureus* gene. The most highly expressed genes are shown in yellow, while genes that are not expressed (RPKM value of 0) are shown in black. A dendrogram representing the hierarchical clustering of each subject is shown below the heatmap.

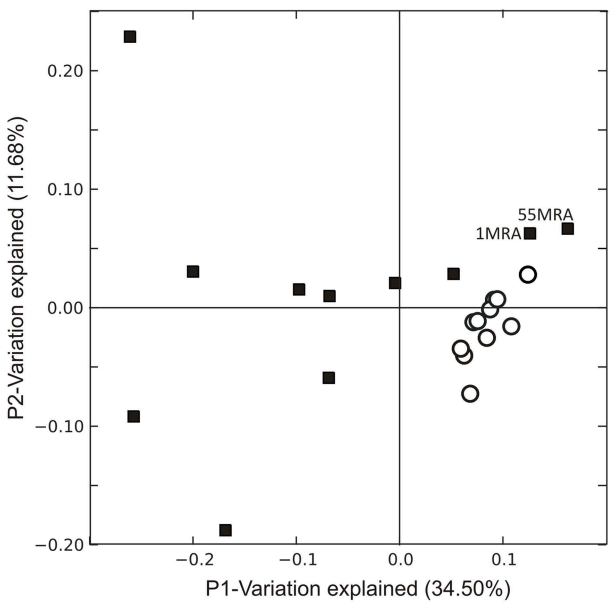


FIGURE 2 | Principal coordinates analysis of *Staphylococcus aureus* transcriptomes *ex vivo* (black squares) and *in vitro* (white circles). Beta diversity was determined using binary Sorensen distances based on RPKM values, and was used as input for the principal coordinates analysis.

1.07% *in vitro*; $p = 0.028$) (**Figure 1A**). As demonstrated by heatmap, many subjects had relatively limited global gene expression profiles *ex vivo* when compared to their *in vitro* counterparts (**Figure 1B**). Samples 21MRA and 23MRA from the same subject demonstrated similar but not identical patterns of gene expression. Overall, MRSA strains had patterns of gene expression that were reflective of either their *ex vivo* or *in vitro* environment (**Figure 1B**), suggesting that the limited patterns of gene expression *ex vivo* directly reflected adaptations to human blood. These data also were supported by principal coordinates analysis, which showed that much less heterogeneity amongst the *in vitro*-derived specimens than was observed for the *ex vivo* specimens (**Figure 2**).

Subsystem Specific Gene Expression in MRSA

We compared MRSA subsystem gene expression profiles to determine whether there were specific differences attributable to culture in media with human blood. There were significant differences ($p = 0.05$) in *ex vivo* and *in vitro* gene expression in numerous subsystems, including virulence/disease/defense, and mobile genetic elements among other subsystems identified (**Supplemental Figure 2**). In 9 of the 10 subjects studied, there was higher *ex vivo* expression of mobile elements than *in vitro* (**Supplemental Figure 3**), and significantly greater expression of prophage genes *in vitro* than *ex vivo* (**Figure 3**). Expression of beta lactamases involved in resistance to beta lactam antibiotics also was significantly increased on transposons (Tn552; $p = 0.007$) and on plasmids ($p = 0.003$). We found no significant differences in expression of the *mecA* penicillin binding protein gene, which also encodes resistance to beta lactam antibiotics.

We also examined the expression of virulence subsystems in MRSA *ex vivo* and *in vitro* to determine the impact of culture in media with human blood on MRSA virulence gene expression. We found that in seven of the 10 subjects studied, virulence gene expression was higher *in vitro* than *ex vivo* (**Supplemental Figure 4**), which suggests that human blood may

have inhibitory effects on the expression of certain virulence genes. Despite the generally lower expression of virulence genes, some were more highly expressed *ex vivo*, including *Staphylococcus aureus* pathogenicity islands (SAPIs) and cytolysins (**Supplemental Figure 5A**). SAPIs are highly diverse phage-related chromosomal islands that insert into the genome at distinct sites (Novick and Subedi, 2007), while cytolysins generally are involved in the lysis of neutrophils (Queck et al., 2009). Only five of the MRSA isolates encoded PVL (lukS/F-PV), which was more highly expressed *ex vivo,* although the difference was not statistically significant ($p = 0.120$) (**Supplemental Figure 5B**).

Because of the high expression of beta lactamases despite the absence of beta lactam therapy in most subjects, we focused on the potential response of MRSA to other antibiotics. Each subject had been treated with vancomycin (a glycopeptide antibiotic that disrupts the cell wall of MRSA through the prevention of cross-linking at terminal d-alanine moieties). We found no significant difference in the expression of *ddl*, which encodes d-alanine moieties (**Supplemental Figure 5B**). *tcaA* also is involved in the response to glycopeptide antibiotics (Srinivasan et al., 2002), and was more highly expressed in all study subjects *ex vivo*, although the difference was not statistically significant ($p = 0.101$).

Prophage Expression Inhibited by Culture in Media with Blood

Because of the significantly higher expression of mobile genetic elements *ex vivo* (**Supplemental Figure 3**), and the greater expression of prophage genes *in vitro* (**Figure 3**), we investigated whether there may be specific differences identified in bacteriophage gene expression *in vitro* and *ex vivo*. There were significant differences in the expression of phage repressors, which were significantly more highly expressed *ex vivo* ($p = 0.039$) (**Figure 4**). Similarly, we also found high expression of phage anti-repressors *in vitro*, although those differences were not statistically significant ($p = 0.238$). The higher expression of repressors and antirepressors *ex vivo* and *in vitro*, respectively, suggests that there were specific interactions with human blood that resulted in inhibition of prophage progression to lytic gene expression.

To test whether culture in media with human blood was inhibitory to prophage expression, we stimulated prophage using mitomycin C in each of the 11 isolated MRSA strains. We found that 6 of the 11 strains produced viable phage in Brain Heart Infusion broth (BHI) in the presence of mitomycin C (including strains 3MRA, 20MRA, 21MRA, 23MRA, 74MRA, and 256MRA); however, the production of phage was inhibited

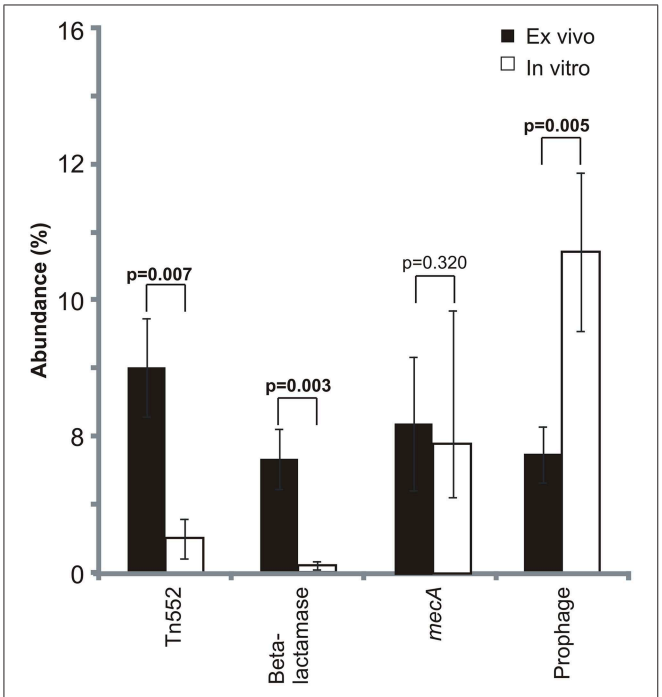


FIGURE 3 | Mean percentage (± standard deviation) of mobile element gene expression devoted to specific elements. Mobile genetic element relative abundances were determined based on the proportion of each selected mobile element to the total mobile genetic element RPKM values. Mobile genetic elements were determined based on homologies to gene categories in the SEED database. *p*-values are shown above each mobile element. *Ex vivo* gene expression is represented by black bars and *in vitro* is represented by white bars.

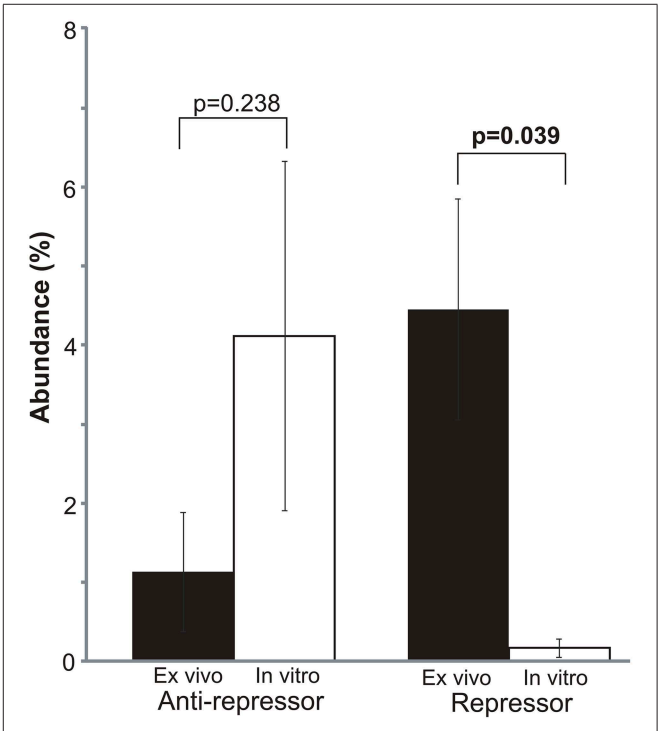


FIGURE 4 | Mean percentage (± standard deviation) of mobile element gene expression devoted to bacteriophage repressors and anti-repressors. Repressor/antirepressor relative abundances were determined based on the proportion of each repressor/antirepressor to the total mobile genetic element RPKM values. Mobile genetic elements were determined based on homologies to gene categories in the SEED database. *p*-values are shown above each element. *Ex vivo* gene expression is represented by black bars and *in vitro* is represented by white bars.

in all six strains by the presence of human blood in culture media (**Figure 5A**). Interestingly, there was spontaneous induction of phage in strains 3MRA and 20MRA, but the number of phage were two to three logs lower than were produced in the presence of mitomycin C. We also tested whether the reduction in phage identified in the presence of blood may be related to inhibition of phage infectivity rather than a decrease in phage production. We found that for both 3MRA and 23MRA there was a decrease in observed PFUs by 1.5 to 2 logs, indicating that infectivity was diminished in the presence of blood (**Figure 5B**). Because phage adsorption has previously been shown to be inhibited by immunoglobulins (Martin and White, 1968; Nordstrom et al., 1974), it is likely that diminished adsorption was responsible for the decrease in infectivity. The fact that infectivity was only partially inhibited (**Figure 5B**) suggests the lack of PFUs produced in culture with blood (**Figure 5A**) cannot be completely explained by inhibition of adsorption. These data are highly suggestive that expression and adsorption of these phage are inhibited by life in the human bloodstream.

Identification of Inhibited Prophage

We sequenced the phage from subjects 3MRA and 23MRA to determine which specific phage were inhibited in culture media with blood. We isolated phage 3MRA and 23MRA starting from single plaques, then purified and sequenced the DNA directly from the virions. We produced a total of 241,857 reads for phage 3MRA, of which 234,536 (97%) assembled into a single 42,141bp contig with 1177X average coverage (**Figure 6A**). We also sequenced 216,080 reads for phage 23MRA, of which 197,211 (91%) assembled into a single 43,114bp contig with 961X average coverage (**Figure 6B**). Based on BLASTN analysis, there were no nearly identical matches for phage 3MRA, but it was similar and shared synteny with *S. aureus* siphoviruses phiETA and phiETA3 (**Figure 7A**). Phage 23MRA was virtually identical to a prophage found in the genome of *S. aureus* USA300 TCH1516, which also is closely related to *S. aureus* siphoviruses phage 77, and phiETA2 (**Figure 7B**). Genome alignments shows that phage 23MRA has much greater similarity to Phage 77 than to phiETA2. Each phage sequenced had homologs to structural genes (head, tail, and portal), replication machinery (polymerases), integration genes (integrases), lysis and packaging machinery (lysins and terminases), genes that control transcription (repressors/anti-repressors), and virulence genes (toxin-antitoxin, complement inhibitor, chemotaxis inhibitor, and staphylokinase). All ORFs identified in phage 3MRA and 23MRA phage had homologs to staphylococcal phage genes or genes previously identified in staphylococcal genomes. These data specifically identify prophage whose expression is inhibited by culture in media with blood, and the profound similarity between these phage and many previously identified staphylococcal prophage suggests that they may also be inhibited by human blood.

Discussion

Staphylococcus aureus is highly prevalent and responsible for substantial morbidity and mortality. Its epidemiology has shifted, resulting in methicillin-resistant strains having similar prevalence to methicillin-sensitive strains in clinical settings. Because studies generally identify single MRSA genotypes from sterile site infections (Young et al., 2004), most MRSA infections are believed to be caused by single rather than multiple different genotypes. There was no evidence based on the MRSA culture data and susceptibility patterns of mixed infections in the 10 subjects studied. While much is known about its pathogenicity, the *in vivo* gene expression of MRSA and its lysogenic phage have yet to be thoroughly examined in humans. Our

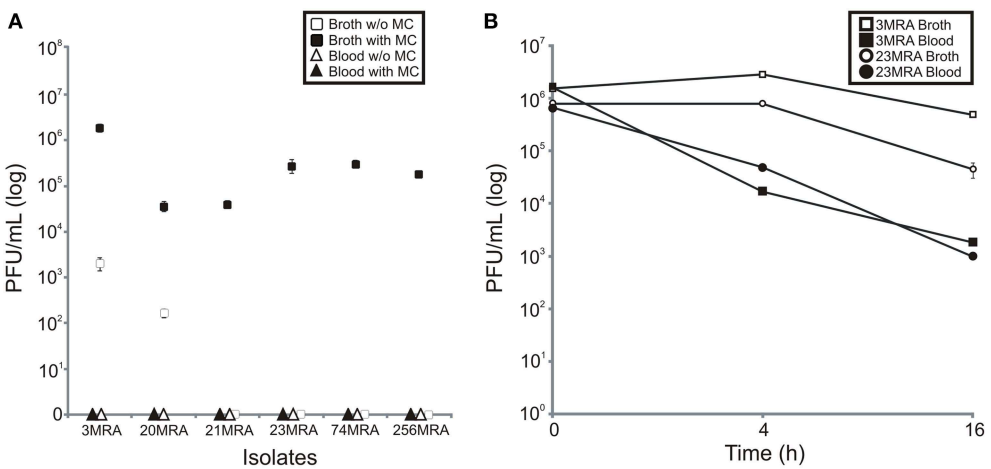


FIGURE 5 | Production of phage from MRSA strains in the presence of human blood, or BHI broth with and without mitomycin C (A), and inhibition of infectivity of phage produced in human blood or BHI broth (B). Mean plaque forming units (log) (± standard deviation) from three separate experiments is demonstrated. In **(A)**, BHI broth with and without mitomycin C are represented by black and white squares, respectively, and blood with and without mitomycin C are represented by black and white triangles, respectively. In **(B)**, phage 3MRA in the presence of BHI broth and blood is represented by white and black squares, respectively, and phage 23MRA in the presence of BHI broth and blood is represented by white and black circles, respectively. Standard deviation bars are demonstrated on each panel.

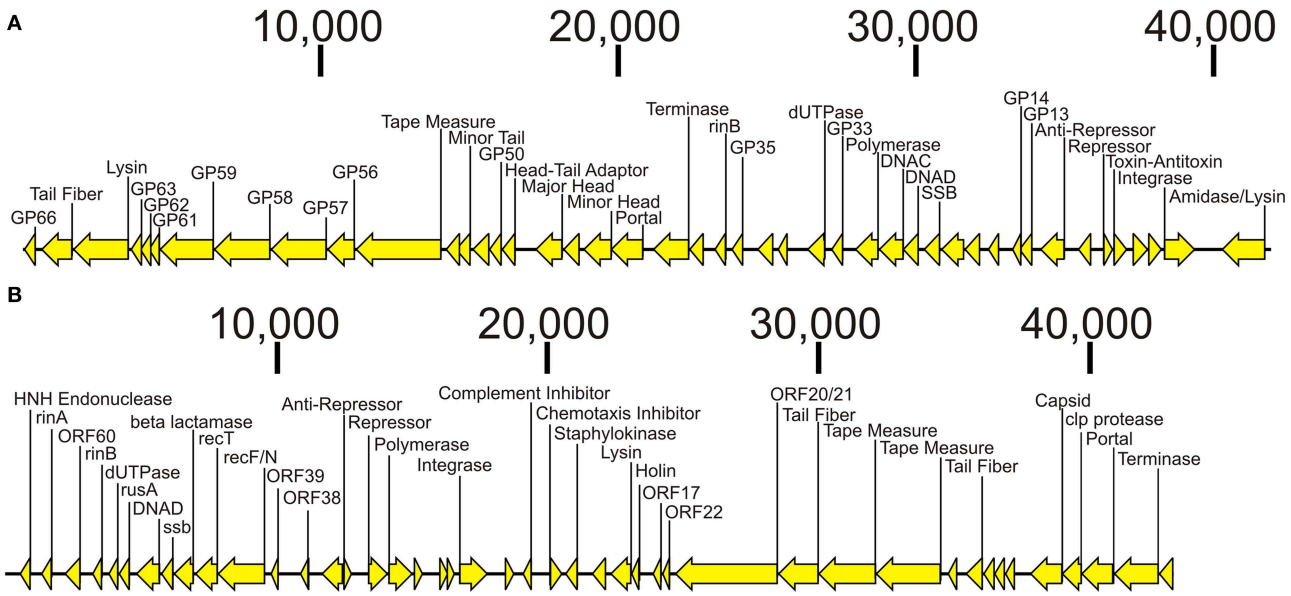


FIGURE 6 | Diagram of bacteriophage genomes sequenced from the virions from Subjects 3MRA (A) and 23MRA (B). Putative ORFs and their direction are indicated by the arrows. ORFs that had significant homologs (BLASTP E-score $<10^{-5}$) are indicated by the text above each arrow. In **(A)**, those ORFs that had synteny with *Staphylococcus* phage PhiETA2 are labeled GP66, GP63, GP62, GP61, GP59, GP58, GP57, GP56, GP50, GP35, GP33, GP14, and GP13. In **(B)**, those ORFs that had synteny with *Staphylococcus* phage PhiPVL are labeled ORF60, ORF39, ORF38, ORF22, ORF21, ORF22, and ORF17. Putative functions for each ORF are labeled when available. The length of each contig is denoted at the top of each panel.

technique for characterizing MRSA *ex vivo* gene expression was unique in that it involved examining the gene expression of MRSA cultured directly from the blood of different human subjects with bloodstream infections without any separate *in vitro* culture steps, followed by comparisons with *in vitro* growth to identify differences that might be attributable to life in the human bloodstream. While not technically an *in vivo* technique because of the separate culture step, we identified numerous differences between *ex vivo* and *in vitro* MRSA gene expression profiles, with the lysogenic phage providing the most notable differences. Although the group of subjects we studied was relatively heterogeneous in their underlying comorbidities, each had invasive MRSA bloodstream infections. The heterogeneity in their comorbidities and individualized medical therapies may be responsible for the variability observed in gene expression from each subject (**Figure 1**). Because it took approximately 72 h to identify those subjects who had MRSA bloodstream infections, most of the subjects enrolled in this study had either cleared their bacteremia or were deceased by the time a diagnosis was made. This severely limited our ability to obtain further specimens to test biological replicates. Because some of the subjects were critically ill and treated with complex antibiotic combinations (**Table 1**), the features of their immune responses and antibiotic regimens would be difficult to replicate *in vitro*, which further adds to the uniqueness of this cohort and the responses of the MRSA isolates studied. The differences in MRSA gene expression profiles observed cannot be solely attributed to antibiotics, as there were different gene expression profiles observed in subjects treated with the same antibiotic regimens.

We were highly interested in the expression of mobile genetic elements (including phage, plasmids, and transposons) associated with bacterial infections in the human bloodstream, as they may be the primary means by which new gene functions are transmitted in nature. Beta lactamases expressed on mobile elements *ex vivo* were significantly overexpressed compared to *in vitro* (**Figure 3**), suggesting that their expression may be part of the organism's stress response to human blood. Because most of the subjects studied were not treated with beta lactam antibiotics, the regulation of these beta lactamases likely is independent of beta lactam therapy. While beta lactamases were more highly expressed *ex vivo*, *mecA* expression was stable under each growth condition. Because the MecA penicillin binding protein is involved in the normal processing of the MRSA cell wall, its expression might not be expected to be substantially altered when cultured in media with human blood.

While there has been some prior analysis of the effect of human blood on prophage (Gaidelyte et al., 2007), our *ex vivo* data strongly suggest that repression of certain prophage is characteristic of MRSA invasive bloodstream infections. This was supported by the higher *ex vivo* expression of repressor genes (**Figure 4**), which are essential for the inhibition of the phage lytic genes. An inverse effect was noted *in vitro*, where the higher expression of anti-repressor genes was indicative of the expression of the phage lytic module. We were unable to place the MRSA genomes into lineage groups (Lindsay et al., 2006; McCarthy and Lindsay, 2010) based on the presence/absence of mobile genetic elements because the heterogeneity in gene expression and the potential for prior horizontal gene transfers (McCarthy et al., 2012a,b) could have led to erroneous results.

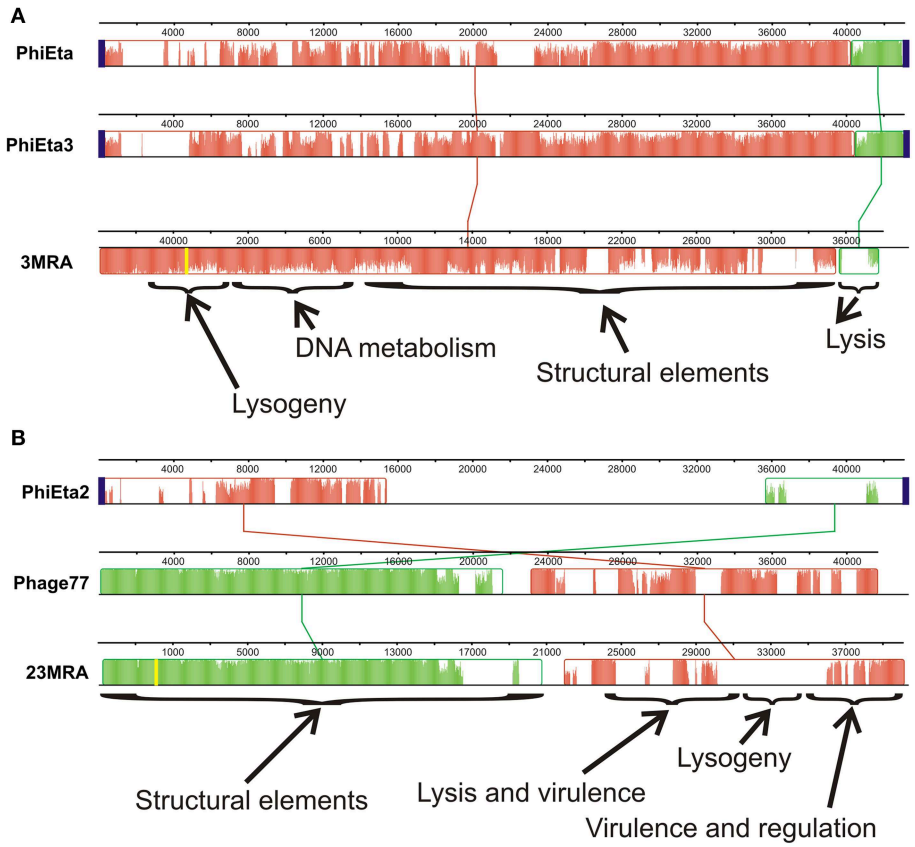


FIGURE 7 | Alignments of phage 3MRA (A) and phage 23MRA (B) with various other known staphylococcal phage. The boxes represent segments of each phage that are relatively well conserved amongst the aligned phage and the lines between the boxes represent the relative position of conserved segments of each phage in the different genomes. Diagonal lines represent potential genome rearrangements. Relative locations in the genomes of each phage are demonstrated by the nucleotide numbers above each phage. The yellow lines represent the sites of the first nucleotide in the assembled 3MRA and 23MRA phage contigs. Some phage segments shown by the arrows contain conserved genes representing structural elements, DNA metabolism, lysis, lysogeny, virulence, and regulation. The height of the colors across each box represents the average conservation in that phage segment across the phage examined. *attP* sites, which represent the sites on phage genomes where they integrate into their host genomes, are represented by blue boxes. *attP* sequences were not determined for phage 3MRA and 23MRA.

There also was altered expression of site-specific recombinases in this cohort, but their potential dual roles in integration and excision complicates their interpretation (Hanssen and Ericson Sollid, 2006; Malachowa and Deleo, 2010).

In addition to altered phage expression in response to blood in culture media, we found that infectivity of the phage also was diminished, likely through inhibition of adsorption (**Figure 5**) (Martin and White, 1968; Nordstrom et al., 1974). These results suggest that the transport and adsorption of phage particles, as well as the transmission of phage-encoded genes, may be restricted during MRSA bloodstream infections in humans. We also identified certain phage that were specifically inhibited by culture in media with human blood (**Figure 6**), and their strong similarities to prophage found in other MRSA isolates (**Figure 7**) suggests that they also may be inhibited by blood. Not all MRSA isolates had produced phage particles that we could detect using acceptor strain RN10950. We also utilized strain RN4220 (a restriction deficient *S. aureus* strain cured of three prophage) (Novick, 1967), but were unable to detect phage from the other MRSA isolates. Our study differs significantly from a prior study documenting the effects of human blood on staphylococcal phage. In that study, the authors concentrated specifically on phage that were expressed during culture in media with human blood (Gaidelyte et al., 2007), whereas our study identifies those phage that were inhibited by culture with human blood. Also, in contrast to that study, our data on these specific MRSA phage shows that infectivity was diminished in the presence of blood (**Figure 5**).

The production of viable phage from prophage in the MRSA genomes in this study directly supported the RNA sequencing data. We observed a substantial over-expression of phage repressors *ex vivo* (**Figure 4**), which correlated directly with the inhibition of phage production in human blood (**Figure 5**). Although most MRSA isolates in our study carried integrases, which often indicate the presence of prophage, not all strains produced viable prophage that we could detect with our acceptor strains. This suggests that some of our MRSA strains carried prophage remnants and/or cryptic prophage rather than viable prophage (Canchaya

et al., 2003; Casjens, 2003). With the exception of isolate 6MRA, most CA-MRSA strains produced viable phage, compared to only one HA-MRSA strain (3MRA). The phage we identified (**Figure 6**) had virulence genes such as staphylokinase (involved in tissue destruction), toxin/antitoxin systems (may be involved in defense, drug tolerance, programmed cell death, and growth control) (Nolle et al., 2013), and genes for an immune escape complex (involved in immune evasion) (Goerke et al., 2006), which suggests that these phage may play several different roles in MRSA pathogenesis.

Conclusions

While some studies have profiled *S. aureus in vitro* behavior (Renzoni et al., 2006; Sass et al., 2008) and gene expression profiles in mouse models (Allard et al., 2006; Chaffin et al., 2012), the gene expression of their bacteriophage and other mobile genetic elements have not been examined *in vivo*. While our *ex vivo* model includes a separate culture step, the blood culturing step was absolutely necessary to first identify those human subjects that had invasive bloodstream infections. The benefit of the model was that we could isolate each strain *in vitro*, and thus discern differences in bacteriophage and mobile genetic elements gene expression profiles between *ex vivo* and *in vitro* that may have represented adaptation to life in the human bloodstream. The expression of mobile genetic elements was amongst the most significant differences between gene expression profiles *ex vivo* and *in vitro*, and these differences identified may have consequences for the human host. From the over-expression of beta lactamases on plasmids and transposons *ex vivo*, to the repression of prophage expression and adsorption during culture in media with human blood, differences in the expression of mobile genetic elements were characteristic of the MRSA *ex vivo* response. As we continue to study the behavior of human pathogens, *ex vivo* studies such as this provide further insights into the gene expression patterns in pathogens and their viruses as they adapt to life in the human host.

Materials and Methods

Ethics Statement

Human subject involvement in this study was approved by the University of California, San Diego Administrative Panel on Human Subjects in Medical Research. The study was certified as category 4 exempt, which includes research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if the information is recorded in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Human Subjects and Culture Conditions

We sampled blood from 10 human subjects with invasive bloodstream infections (**Table 1**). Samples 21MRA and 23MRA were taken from the same subject approximately 48 h apart for a total of 11 different samples from 10 subjects. From each subject, 8–10 mL of blood was inoculated directly at bedside into 30 ml of media in BD Bactec Plus aerobic culture vials (BD Diagnostic Systems, Sparks, MD) and incubated for approximately 4–8 h until positive by fluorescent detection using the Bactec FX system (Zadroga et al., 2013). Positive cultures were gram stained, grown on sheep blood agar plates, and subjected to a coagulase test for a presumptive identification of *Staphylococcus aureus*. Identification and susceptibility testing was performed using the BD Phoenix system (Stefaniuk et al., 2003) to confirm the presence of MRSA in the bloodstream of each subject. Susceptibilities were based on MIC (minimum inhibitory concentration) breakpoints using Clinical and Laboratory Standards Institute guidelines (CLSI, 2012). Each isolate was assayed for susceptibility to antibiotics cefazolin, clindamycin, daptomycin, linezolid, oxacillin, rifampin, trimethoprim/sulfamethoxazole, and vancomycin. Clindamycin susceptibility was verified by D-test (Woods, 2009). A minimum of 3 mL of each positive sample in the standard aerobic culture vial was immediately processed in the *ex vivo* arm of this study. For *the in vitro* growth conditions, each identified MRSA strain was grown on blood agar plates, and reconstituted in tryptic soy broth at an OD600 of 0.1. One milliliter of this suspension was diluted in 7 ml of sterile normal saline and inoculated into the BD Bactec Plus aerobic culture vials and incubated for approximately 3–6 h until positive by fluorescent detection using the Bactec FX system. One milliliter of this suspension was immediately processed in the *in vitro* arm of this study.

MRSA Growth Conditions

OD_{600} values for each isolate grown in the BD Bactec Plus aerobic vials was determined at the time of fluorescence detection (approximately 3.5 to 4.5 h, depending on the isolate). To determine the growth phase at the time of fluorescence detection, MRSA isolates were grown for 16 h in BD Bactec Plus aerobic broth and reconstituted to an OD_{600} of 0.1. Cultures were then incubated at 35°C in BD Bactec Plus aerobic broth with gentle agitation and OD_{600} values for each isolate was determined at 15 min intervals over a 24 h period to construct growth curves. CFU/mL for each isolate under *in vitro* conditions were determined by plating 100 μL of serial dilutions onto BHI agar plates and incubated at 35°C for 16 h. Because the presence of the blood prohibited us from using direct comparisons of OD_{600} values between the *ex vivo* and *in vitro* cultures, we determined the CFU/mL counts from each subject *ex vivo* to estimate the growth phase at the time of fluorescence detection by plating 100 μL of serial dilutions on BHI agar plates. All *ex vivo* and *in vitro* cultures were found to be in early to mid log phase at the time of detection. Two volumes of RNA protect (Qiagen, Valencia, CA) was added directly to the cultures, and they were pelleted and stored at −20°C until RNA extraction.

RNA Extraction, Enrichment, and Sequencing

RNA from both the *ex vivo* and *in vitro* samples were processed identically from all subjects and MRSA isolates. Total RNA was extracted using the Mirvana kit (Life Technologies, Grand Island, NY), with the inclusion of a bead-beating step for 20 min with Lysing-Matrix B (MP Bio, Santa Ana, CA). Total RNA then was enriched for microbial RNA using

MicrobEnrich (Life Technologies), and further enriched for mRNA using MicrobExpress (Life Technologies) and MegaClear (Life Technologies), which are designed to remove ribosomal RNAs. Enriched RNA then was prepared for sequencing through the construction of cDNA libraries using the Ion Total RNA-Seq kit (Life Technologies), and subjected to successive rounds of Ampure bead purification (Beckman-Coulter, Brea, CA) to remove small cDNAs. Libraries were quantified using an Agilent Bioanalyzer HS DNA Kit (Agilent, Santa Clara, CA) and then were sequenced on a 314 chips using an Ion Torrent Personal Genome Machine (Rothberg et al., 2011), producing an average of 559,129 reads per subject of mean length 106 nucleotides. All sequence data produced in this study are available in the MG-Rast database (metagenomics.anl.gov/) under the project name "MRSA_RNAseq_Study" or project #2278.

Processing of RNA Sequences

Sequencing reads were trimmed according to modified Phred quality scores of 0.5 using CLC Genomics Workbench 4.65 (CLC bio USA, Cambridge, MA). The remaining reads were further processed for quality control by removing reads with substantial length variation (reads <50 nucleotides or >200 nucleotides), or reads where $\geq 25\%$ of the length was due to homopolymers tracts. Each transcriptome was screened for contaminating human nucleic acids using BLASTN analysis (E-score $<10^{-5}$) against the human reference database available at ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/. Any reads homologous to human sequences were removed prior to further analysis. Both *in vitro* and *ex vivo* reads from each subject were mapped to a database of staphylococcal genomes (available at ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) to determine the percentage of reads that mapped to *Staphylococcus* and to which individual strains they mapped best. Reads from each subject mapped best to known MRSA genomes in all subjects under each growth condition (**Table 2**).

Analysis of Transcriptomes

Each *ex vivo* and *in vitro* transcriptome were mapped against USA300 MRSA strain FPR3757 using CLC Genomics Workbench 4.65. RPKM (reads per kilobase per million) values were determined based on the read mappings, and values were normalized for each subject and growth condition. A heatmap demonstrating the distribution of gene expression based on normalized RPKM values was generated using CLC Genomics Workbench 4.65. Virtually identical heatmaps also were generated when reads were mapped against genomes of other MRSA strains including COL, JH1, Mu50, TCH1516, and Newman. RPKM values also were used as input for principal coordinates analysis, and were performed based on binary Sorensen distances using Qiime (Caporaso et al., 2010).

Analysis of gene expression from different subsystems in MRSA were determined by blastX analysis of the SEED database using MG-Rast (E-score $<10^{-5}$) (Meyer et al., 2008). Mobile genetic element and virulence gene expression was determined based on the proportion of each selected mobile element or virulence gene to the total mobile genetic element or virulence gene RPKM values. Subsystems such as mobile genetic elements and virulence genes were determined based on homologies to genes designated to those functions in the SEED database (Meyer et al., 2008). Statistical significance was determined by comparing the means for all subjects for all subsystems between the *ex vivo* and *in vitro* subject groups by two-tailed *t*-tests using Microsoft Excel 2007 (Microsoft Corp., Redman, WA). Analysis of the expression differences between *in vitro* and *ex vivo* groups for individual genes also were determined by comparisons of means. The data for each individual gene were compared between MG-Rast and the normalized RPKM values obtained from CLC Genomics Workbench 4.65 to verify that they produced similar results.

Prophage Stimulation and Sequencing

Each MRSA strain was grown in BHI broth (Becton, Dickinson, MD, USA) for approximately 3 h at 37°C with shaking at 200 rpm to an OD_{600} of 0.5. The average number of cells was similar between each MRSA strain ($1.84 \pm 0.45 \times 10^8$ Colony Forming Units (CFU)/mL). Each strain then was centrifuged for 10 min at 10,000 rpm, pellets re-suspended in 500 μL of BHI broth or human blood (Novick, 1963), and mitomycin C added to a final concentration of 2 μg/mL. The human blood used was drawn into heparin tubes (BD Diagnostic Systems, Sparks, MD) from healthy donors. MRSA cultures without the addition of mitomycin C were used as experimental controls and were without evidence of any prophage production, with the exception of isolates 3MRA and 20MRA. Each was left for 16 h at 32°C with shaking at 50 rpm. To determine the number of phage produced, cultures were centrifuged for 5 min at 14, 000 rpm to remove cellular debris, and filtered through 0.2 μm pore membrane filters (25 mm, Whatman GE Healthcare Life Sciences). Supernatants were cultured for the presence of any residual bacteria, and no bacteria could be cultured from any. We tested the supernatants for the presence of viable phage using the double layer method (Novick, 1991). Briefly, *S. aureus* strain RN10950 (*S. aureus* Newman strain with all four prophage deleted) (Bae et al., 2006) was grown for 4 h at 37°C with shaking at 200 rpm to an OD_{600} of 1.0. Then, 100 μL of *S. aureus* RN10950 and 100 μL of the phage supernatants were added to 3 mL of phage top agar and poured onto phage bottom agar plates (Novick, 1991). Plates were incubated at 32°C for 16 h and viral plaques were enumerated and reported as Plaque Forming Units (PFU)/mL.

To determine if blood may inhibit phage infectivity, phage 3MRA and 23MRA were added to 1 mL of BHI or fresh human blood and incubated at 32°C with shaking at 50 rpm to maintain the conditions used in induction experiments. Aliquots were collected at 0, 4, and 16 h to determine the infectivity of phage. Blood cells were removed prior to plating phage by centrifugation at 14,000 rpm for 5 min and filtered through 0.45 μm filters. The phage in BHI also was centrifuged and filtered for consistency. *S. aureus* strain RN10950 was grown for 4 h at 37°C with shaking at 200 rpm to an OD600 of 1.0. Then, 100 μL of RN10950 and 100 μL of BHI or blood containing the seeded phage were serially diluted and added to 3 mL of phage top agar and poured onto phage bottom agar plates. Plates were incubated at 32°C for 16 h and viral plaques were enumerated and reported as PFU/mL. All experiments were performed in triplicate.

Phage were isolated as previously described (Santiago-Rodriguez et al., 2010). Briefly, viral plaques were retrieved using a sterile pipette and placed in 500 μL of 1× PBS. The plug was dislodged using a sterile pipette and centrifuged at 14,000 rpm for 5 min. The supernatant was collected and further propagated by adding 100 μL of the phage supernatant and 100 μL (OD_{600} of 1.0) of *S. aureus* RN10950 to 3 mL of top agar. The mixture was poured onto phage bottom agar plates and incubated at 32°C for 24 h. The top agar was collected and centrifuged at 7500 rpm for 15 min. The supernatant was collected and filtered sequentially through 0.45 μm filters and 0.22 μm filters. The propagation step was repeated until complete lysis was observed. Supernatants then were purified on a cesium chloride gradient according to previously described protocols for isolation of viruses (Pride et al., 2012). Only the fraction with a density corresponding to most known bacteriophage (Murphy et al., 1995) was retained, further purified on Amicon YM-100 protein purification columns (Millipore, Inc., Bellerica, MA), treated with DNase I, and subjected to lysis and DNA purification using the Qiagen UltraSens virus kit (Qiagen, Valencia, CA). Resulting DNA was fragmented to roughly 200 to 400 bp using a Bioruptor (Diagenode, Denville, NJ), and utilized as input to create libraries using the Ion Plus Fragment Library Kit according to manufacturer's instructions. Libraries then were sequenced using a 314 chip on an Ion Torrent Personal Genome Machine (PGM; Life Technologies, Grand Island, NY) (Rothberg et al., 2011) producing an average read length of approximately 216 bp for each sample. Sequence reads were trimmed according to modified Phred scores of 0.5 using CLC Genomics Workbench 4.65 (CLC bio USA, Cambridge, MA). Any low complexity reads (where >25% of the length were due to homopolymer tracts), reads with substantial length variation (<50 nucleotides or >300 nucleotides), or reads with ambiguous characters also were removed prior to further analysis. Remaining reads were assembled using CLC Genomics Workbench 4.65 based on 98% identity with a minimum of 50% read overlap, which are more stringent than criteria developed to discriminate between highly related viruses (Breitbart et al., 2002). The consensus sequence for each assembled phage was constructed according to majority rule. Viral contigs were analyzed using FGenesV (Softberry Inc, Mount Kisco, NY) for ORF prediction, and individual ORFs analyzed using BLASTP analysis against the NCBI non-redundant database (Escore $<10^{-5}$). Alignments of phage genomes were performed with progressiveMauve using the default settings (Darling et al., 2010). Sequences of each *Staphylococcus* phage are available in Genbank under accession numbers KJ452291 and KJ452292. The reverse complements of these phage genome sequences were utilized in the genome alignments.

Author Contributions

Conceived and designed experiments: DTP and MBJ. Performed the experiments TSR, MN, and ML. Analyzed the data: DTP, TSR, and MBJ. Wrote the manuscript: DTP and TSR. All authors have read and approved this manuscript.

Acknowledgments

Supported by the Burroughs Wellcome Fund, and the UNCF-Merck Science Initiative to DTP. We thank Sharon Reed, Lizanne Keays, Jane Harrington, and Lars Eckman for their contribution to this work. Strain RN10950 was donated by Dr. Richard Novick.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00216/abstract

Supplemental Figure 1 | Growth curves for MRSA isolates. Each isolate is shown in separate panels with their respective OD_{600} measurements at 15 min intervals on the y-axis and the time over a period of 24 h period shown on the x-axis. The time of fluorescence detection under *in vitro* (blue circles) and *ex vivo* (red circles) conditions are shown for each subject in each panel.

Supplemental Figure 2 | Mean percentage (± standard deviation) of all gene expression from each subsystem in *Staphylococcus aureus* transcriptomes *ex vivo* and *in vitro*. The upper panel shows subsystems that were significantly different between *ex vivo* and *in vitro* samples, and the lower panel shows subsystems that were not significantly different. *p*-values are shown to the right of each subsystem. *Ex vivo* gene expression is represented by black bars and *in vitro* is represented by white bars.

Supplemental Figure 3 | Abundance (%) of mobile element gene expression in *Staphylococcus aureus* transcriptomes for all subjects. Specimens 21MRA and 23MRA are from the same subject 48 h apart.

Supplemental Figure 4 | Abundance (%) of virulence gene expression in *Staphylococcus aureus* transcriptomes for all subjects. *Ex vivo* gene expression is represented by black bars and *in vitro* is represented by white bars. Specimens 21MRA and 23MRA are from the same subject 48 h apart.

Supplemental Figure 5 | Mean percentage (± standard deviation) of the virulence subsystem devoted to individual elements. *p*-values are shown above each virulence factor. *Ex vivo* gene expression is represented by black bars and *in vitro* is represented by white bars. **(A)** shows Staphylococcal Pathogenicity Islands (SaPIs) and cytolysins, and **(B)** shows *tcaA* (teicoplanin-resistance), *ddl* (involved in cell wall synthesis) and *lukS-PV* and *lukF-PV* (encoding Panton-Valentine leukocidin).

References

Allard, M., Moisan, H., Brouillette, E., Gervais, A. L., Jacques, M., Lacasse, P., et al. (2006). Transcriptional modulation of some *Staphylococcus aureus* iron-regulated genes during growth *in vitro* and in a tissue cage model *in vivo*. *Microbes Infect.* 8, 1679–1690. doi: 10.1016/j.micinf.2006.01.022

Baba, T., Bae, T., Schneewind, O., Takeuchi, F., and Hiramatsu, K. (2008). Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. *J. Bacteriol.* 190, 300–310. doi: 10.1128/JB.01000-07

Bae, T., Baba, T., Hiramatsu, K., and Schneewind, O. (2006). Prophages of *Staphylococcus aureus* Newman and their contribution to virulence. *Mol. Microbiol.* 62, 1035–1047. doi: 10.1111/j.1365-2958.2006.05441.x

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Brussow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602. doi: 10.1128/MMBR.68.3.560-602.2004

Canchaya, C., Proux, C., Fournous, G., Bruttin, A., and Brussow, H. (2003). Prophage genomics. *Microbiol. Mol. Biol. Rev.* 67, 238–276. doi: 10.1128/MMBR.67.2.238-276.2003

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Meth.* 7, 335–336. doi: 10.1038/nmeth.f.303

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Mol. Microbiol.* 49, 277–300. doi: 10.1046/j.1365-2958.2003.03580.x

Chaffin, D. O., Taylor, D., Skerrett, S. J., and Rubens, C. E. (2012). Changes in the *Staphylococcus aureus* transcriptome during early adaptation to the lung. *PLoS ONE* 7:e41329. doi: 10.1371/journal.pone.0041329

CLSI, (2012). "Performance Standards for Antimicrobial Susceptibility Testing; Twenty-Second Informational Supplement," in *CLSI Document M100-S22*, ed D. M. Wilhelm (Wayne, PA: Clinical and Laboratory Standards Institute), 110–115.

Cui, L., Lian, J. Q., Neoh, H. M., Reyes, E., and Hiramatsu, K. (2005). DNA microarray-based identification of genes associated with glycopeptide resistance in *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* 49, 3404–3413. doi: 10.1128/AAC.49.8.3404-3413.2005

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147. doi: 10.1371/journal.pone.0011147

Deghorain, M., and Van Melderen, L. (2012). The Staphylococci phages family: an overview. *Viruses* 4, 3316–3335. doi: 10.3390/v4123316

Diep, B. A., Palazzolo-Ballance, A. M., Tattevin, P., Basuino, L., Braughton, K. R., Whitney, A. R., et al. (2008a). Contribution of Panton-Valentine leukocidin in community-associated methicillin-resistant *Staphylococcus aureus* pathogenesis. *PLoS ONE* 3:e3198. doi: 10.1371/journal.pone.0003198

Diep, B. A., Stone, G. G., Basuino, L., Graber, C. J., Miller, A., Des Etages, S. A., et al. (2008b). The arginine catabolic mobile element and staphylococcal chromosomal cassette mec linkage: convergence of virulence and resistance in the USA300 clone of methicillin-resistant *Staphylococcus aureus*. *J. Infect. Dis.* 197, 1523–1530. doi: 10.1086/587907

Feng, Y., Chen, C. J., Su, L. H., Hu, S., Yu, J., and Chiu, C. H. (2008). Evolution and pathogenesis of *Staphylococcus aureus*: lessons learned from genotyping and comparative genomics. *FEMS Microbiol. Rev.* 32, 23–37. doi: 10.1111/j.1574-6976.2007.00086.x

Gaidelyte, A., Vaara, M., and Bamford, D. H. (2007). Bacteria, phages and septicemia. *PLoS ONE* 2:e1145. doi: 10.1371/journal.pone.0001145

Goerke, C., Wirtz, C., Fluckiger, U., and Wolz, C. (2006). Extensive phage dynamics in *Staphylococcus aureus* contributes to adaptation to the human host during infection. *Mol. Microbiol.* 61, 1673–1685. doi: 10.1111/j.1365-2958.2006.05354.x

Hanssen, A. M., and Ericson Sollid, J. U. (2006). SCCmec in staphylococci: genes on the move. *FEMS Immunol. Med. Microbiol.* 46, 8–20. doi: 10.1111/j.1574-695X.2005.00009.x

Kennedy, L. A., Gill, J. A., Schultz, M. E., Irmler, M., and Gordin, F. M. (2010). Inside-out: the changing epidemiology of methicillin-resistant *Staphylococcus aureus*. *Infect. Control Hosp. Epidemiol.* 31, 983–985. doi: 10.1086/655837

Lindsay, J. A., Moore, C. E., Day, N. P., Peacock, S. J., Witney, A. A., Stabler, R. A., et al. (2006). Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J. Bacteriol.* 188, 669–676. doi: 10.1128/JB.188.2.669-676.2006

Malachowa, N., and Deleo, F. R. (2010). Mobile genetic elements of *Staphylococcus aureus*. *Cell Mol. Life Sci.* 67, 3057–3071. doi: 10.1007/s00018-010-0389-4

Malachowa, N., and Deleo, F. R. (2011). *Staphylococcus aureus* survival in human blood. *Virulence* 2, 567–569. doi: 10.4161/viru.2.6.17732

Malachowa, N., Whitney, A. R., Kobayashi, S. D., Sturdevant, D. E., Kennedy, A. D., Braughton, K. R., et al. (2011). Global changes in *Staphylococcus aureus* gene expression in human blood. *PLoS ONE* 6:e18617. doi: 10.1371/journal.pone.0018617

Martin, R. R., and White, A. (1968). Prevention of staphylococcal bacteriophage activity by antigen A precipitins in human sera. *J. Bacteriol.* 95, 2177–2181.

McCarthy, A. J., Breathnach, A. S., and Lindsay, J. A. (2012a). Detection of mobile-genetic-element variation between colonizing and infecting hospital-associated methicillin-resistant *Staphylococcus aureus* isolates. *J. Clin. Microbiol.* 50, 1073–1075. doi: 10.1128/JCM.05938-11

McCarthy, A. J., and Lindsay, J. A. (2010). Genetic variation in *Staphylococcus aureus* surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. *BMC Microbiol.* 10:173. doi: 10.1186/1471-2180-10-173

McCarthy, A. J., Witney, A. A., and Lindsay, J. A. (2012b). *Staphylococcus aureus* temperate bacteriophage: carriage and horizontal gene transfer is lineage associated. *Front. Cell Infect. Microbiol.* 2:6. doi: 10.3389/fcimb.2012.00006

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Murphy, F. A., Fauquet, C. M., Bishop, D. H. L., Ghabrial, S. A., Jarvis, A. W., et al. (1995). *Virus Taxonomy: Sizth Report of the International Committee on Taxonomy of Viruses.* Vol. Suppl. 10. New York, NY: Springer-Verlag.

Nolle, N., Schuster, C. F., and Bertram, R. (2013). Two paralogous yefM-yoeB loci from Staphylococcus equorum encode functional toxin-antitoxin systems. *Microbiology* 159, 1575–1585. doi: 10.1099/mic.0.068049-0

Nordstrom, K., Forsgren, A., and Cox, P. (1974). Prevention of bacteriophage adsorption to *Staphylococcus aureus* by immunoglobulin G. *J. Virol.* 14, 203–206.

Novick, R. (1967). Properties of a cryptic high-frequency transducing phage in *Staphylococcus aureus*. *Virology* 33, 155–166. doi: 10.1016/0042-6822(67)90105-5

Novick, R. P., and Subedi, A. (2007). The SaPIs: mobile pathogenicity islands of Staphylococcus. *Chem. Immunol. Allergy* 93, 42–57. doi: 10.1159/000100857

Novick, R. P. (1963). Analysis by transduction of mutations affecting Penicillinase formation in *Staphylococcus Aureus*. *J. Gen. Microbiol.* 33, 121–136. doi: 10.1099/00221287-33-1-121

Novick, R. P. (1991). Genetic systems in staphylococci. *Methods Enzymol.* 204, 587–636. doi: 10.1016/0076-6879(91)04029-N

Pan, E. S., Diep, B. A., Carleton, H. A., Charlebois, E. D., Sensabaugh, G. F., Haller, B. L., et al. (2003). Increasing prevalence of methicillin-resistant *Staphylococcus aureus* infection in California jails. *Clin. Infect. Dis.* 15, 1384–1388. doi: 10.1086/379019

Pohl, K., Francois, P., Stenz, L., Schlink, F., Geiger, T., Herbert, S., et al. (2009). CodY in *Staphylococcus aureus*: a regulatory link between metabolism and virulence gene expression. *J. Bacteriol.* 191, 2953–2963. doi: 10.1128/JB.01492-08

Popovich, K. J., Weinstein, R. A., and Hota, B. (2008). Are community-associated methicillin-resistant *Staphylococcus aureus* (MRSA) strains replacing traditional nosocomial MRSA strains? *Clin. Infect. Dis.* 46, 787–794. doi: 10.1086/528716

Pride, D. T., Salzman, J., Haynes, M., Rohwer, F., Davis-Long, C., White, R. A., et al. (2012). Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J.* 6, 915–926. doi: 10.1038/ismej.2011.169

Queck, S. Y., Khan, B. A., Wang, R., Bach, T. H., Kretschmer, D., Chen, L., et al. (2009). Mobile genetic element-encoded cytolysin connects virulence to methicillin resistance in MRSA. *PLoS Pathog* 5:e1000533. doi: 10.1371/journal.ppat.1000533

Renzoni, A., Barras, C., Francois, P., Charbonnier, Y., Huggler, E., Garzoni, C., et al. (2006). Transcriptomic and functional analysis of an autolysis-deficient, teicoplanin-resistant derivative of methicillin-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* 50, 3048–3061. doi: 10.1128/AAC.00113-06

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475, 348–352. doi: 10.1038/nature10242

Sabersheikh, S., and Saunders, N. A. (2004). Quantification of virulence-associated gene transcripts in epidemic methicillin resistant *Staphylococcus aureus* by real-time PCR. *Mol. Cell Probes.* 18, 23–31. doi: 10.1016/j.mcp.2003.07.009

Santiago-Rodriguez, T. M., Davila, C., Gonzalez, J., Bonilla, N., Marcos, P., Urdaneta, M., et al. (2010). Characterization of Enterococcus faecalis-infecting phages (enterophages) as markers of human fecal pollution in recreational waters. *Water Res.* 44, 4716–4725. doi: 10.1016/j.watres.2010.07.078

Sass, P., Jansen, A., Szekat, C., Sass, V., Sahl, H. G., and Bierbaum, G. (2008). The lantibiotic mersacidin is a strong inducer of the cell wall stress response of *Staphylococcus aureus*. *BMC Microbiol.* 8:186. doi: 10.1186/1471-2180-8-186

Shopsin, B., Herring, S., and Kreiswirth, B. N. (2003). Hospital-acquired and community-derived: the future of MRSA? *Clin. Infect. Dis.* 37, 151–152; author reply 152. doi: 10.1086/375608

Srinivasan, A., Dick, J. D., and Perl, T. M. (2002). Vancomycin resistance in staphylococci. *Clin. Microbiol. Rev.* 15, 430–438. doi: 10.1128/CMR.15.3.430-438.2002

Stefaniuk, E., Baraniak, A., Gniadkowski, M., and Hryniewicz, W. (2003). Evaluation of the BD Phoenix automated identification and susceptibility testing system in clinical microbiology laboratory practice. *Eur. J. Clin. Microbiol. Infect. Dis.* 22, 479–485. doi: 10.1007/s10096-003-0962-y

Stevens, D. L., Ma, Y., Salmi, D. B., McIndoo, E., Wallace, R. J., and Bryant, A. E. (2007). Impact of antibiotics on expression of virulence-associated exotoxin genes in methicillin-sensitive and methicillin-resistant *Staphylococcus aureus*. *J. Infect. Dis.* 195, 202–211. doi: 10.1086/510396

Vandenesch, F., Naimi, T., Enright, M. C., Lina, G., Nimmo, G. R., Heffernan, H., et al. (2003). Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton-Valentine leukocidin genes: worldwide emergence. *Emerg. Infect. Dis.* 9, 978–984. doi: 10.3201/eid0908.030089

Wilhelm, B. T., and Landry, J. R. (2009). RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 48, 249–257. doi: 10.1016/j.ymeth.2009.03.016

Witney, A. A., Marsden, G. L., Holden, M. T., Stabler, R. A., Husain, S. E., Vass, J. K., et al. (2005). Design, validation, and application of a seven-strain *Staphylococcus aureus* PCR product microarray for comparative genomics. *Appl. Environ. Microbiol.* 71, 7504–7514. doi: 10.1128/AEM.71.11.7504-7514.2005

Woods, C. R. (2009). Macrolide-inducible resistance to clindamycin and the D-test. *Pediatr. Infect. Dis. J.* 28, 1115–1118. doi: 10.1097/INF.0b013e3181c35cc5

Young, L. S., Perdreau-Remington, F., and Winston, L. G. (2004). Clinical, epidemiologic, and molecular evaluation of a clonal outbreak of methicillin-resistant *Staphylococcus aureus* infection. *Clin. Infect. Dis.* 38, 1075–1083. doi: 10.1086/382361

Zadroga, R., Williams, D. N., Gottschall, R., Hanson, K., Nordberg, V., Deike, M., et al. (2013). Comparison of 2 blood culture media shows significant differences in bacterial recovery for patients on antimicrobial therapy. *Clin. Infect. Dis.* 56, 790–797. doi: 10.1093/cid/cis1021

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

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 January 2015
doi: 10.3389/fmicb.2015.00003

Replicating phages in the epidermal mucosa of the eel (*Anguilla anguilla*)

Miguel Carda-Diéguez[1], Carolina Megumi Mizuno[2], Rohit Ghai[2], Francisco Rodriguez-Valera[2] and Carmen Amaro[1]*

[1] *ERI Biotecmed, University of Valencia, Valencia, Spain*
[2] *Evolutionary Genomics Group, Department de Producción Vegetal y Microbiología, Universidad Miguel Hernández, San Juan de Alicante, Spain*

Edited by:
Bas E. Dutilh, Radboud University Medical Center, Netherlands

Reviewed by:
Grzegorz Wegrzyn, University of Gdansk, Poland
Jeremy J. Barr, San Diego State University, USA

****Correspondence:***
Carmen Amaro, ERI Biotecmed, University of Valencia, Doctor Moliner 50, Burjassot 46100, Valencia, Spain
e-mail: carmen.amaro@uv.es

In this work, we used the eel (*Anguilla anguilla*) as an animal model to test the hypothesis of Barr et al. (2013a,b) about the putative role of the epidermal mucosa as a phage enrichment layer. To this end, we analyzed the microbial content of the skin mucus of wild and farmed eels by using a metagenomic approach. We found a great abundance of replicating phage genomes (concatemers) in all the samples. They were assembled in four complete genomes of three Myovirus and one Podovirus. We also found evidences that ΦKZ and Podovirus phages could be part of the resident microbiota associated to the eel mucosal surface and persist on them over the time. Moreover, the viral abundance estimated by epiflorescent counts and by metagenomic recruitment from eel mucosa was higher than that of the surrounding water. Taken together, our results support the hypothesis that claims a possible role of phages in the animal mucus as agents controlling bacterial populations, including pathogenic species, providing a kind of innate immunity.

Keywords: metagenomics, phage, eel, mucosa, immunity

INTRODUCTION

Animals, including humans, protect their mucous membranes by covering them with a mucus layer that contains microcidal and microstatic compounds produced by epidermis-associated immune cells. Being permanently immersed, fish mucus attracts aquatic bacteria, including pathogens, some of which have chemotactic systems that "sense" mucin, glucids, and other nutrients present in mucus (Valiente et al., 2008). For this reason, fish exposed surfaces are covered by a dense mucus layer enriched in antimicrobial compounds, most of them uncharacterized (Ellis, 2001). While most aquatic bacteria are supposed to be controlled by mucus antimicrobial compounds, pathogens have developed mechanisms to resist their attack and colonize mucosal surfaces.

Recently, Barr et al. (2013a,b) proposed a hypothesis suggesting an additional protective function for the mucus: to trap bacteriophages in order to detect and destroy invading bacteria before they reach to the epithelium. The attraction of virus to mucus surfaces has been corroborated in coral mucus using enumeration techniques (Nguyen-Kim et al., 2014). If this hypothesis is true, phages could play a significant role in defense against aquatic pathogens. However, the study of phages present in environmental samples is not easy. Traditional approaches are very limited as they rely on previous knowledge about the host and its culture. On the other hand, metaviromic approaches are constrained by the low amount of viral DNA in natural samples that forces the use of heavily biased amplification steps (Minot et al., 2011; Phan et al., 2011; Hurwitz and Sullivan, 2013; Zablocki et al., 2014). However, metagenomic DNA (derived from the cellular fraction) is known to contain a high proportion of phage DNA. Caudovirales lytic phages replicate forming genome concatamers that can be larger than the cell genome and provide an ideal subject for direct sequencing and assembly (Mizuno et al., 2013; Rodriguez-Valera et al., 2014).

The objective of the present study was to test the hypothesis of Barr et al. (2013a,b) by using the European eel (*Anguilla anguilla*) as a model fish and the metagenomic approach mentioned above to identify the viruses present in the mucus. Eels are quite resistant to stressors and are well equipped against pathogens (Ellis, 2001). Their lack of protective macroscopic scales is compensated with a thick mucus layer especially rich in antimicrobial compounds (Tesch, 2003). We postulate that this protection is partially due to the mucus-attracted phages. We have sampled the mucus from wild eels captured in three Mediterranean wet lands (Albufera Lake, Ebro Delta, and Cabanes; Figure S1). In addition, the European eel is also a commercially important species in Europe, Asia, New Zealand, and USA, that is produced in farms, although its life cycle has not been closed and eel farmers fatten juvenile glass eels caught from natural stocks (Lee et al., 2003). Therefore, we have used also farmed eels, and used a metagenomic approach to identify replicating phages. We have analyzed the water metagenome from one of the sampling sites (Ebro Delta) for comparison. We describe the genomes of four complete, abundant phages (three Myoviruses and one Podovirus), plus one incomplete myovirus, that were directly assembled from the metagenomic reads. This is the first report of abundant phages in the epidermal mucosa of eels or any other fish species.

MATERIALS AND METHODS

SAMPLING

Three natural parks in the Mediterranean Spanish coast were chosen to fish wild eels: Albufera de Valencia (Valencia, 39°19′54″N,

0°21′8″W), Alfacada pond (40° 40′ 45.609″N, 0° 49′ 52.7982″W) in the Ebro delta, two samples a year apart (2013–2014), and the Prado Cabanes (Castellón 40°12′03.5″N 0°12′26.4″E; Figure S1). Moreover, a volume of 25 L of water from Ebro delta (2014) was also analyzed.

To recover eel mucus the animals were kept in a fish bowl with 1 L of PBS (1% NaCl) during 15–20 min. Previous experience has shown that 50–100 ml of mucus is released by this procedure by 5–10 adult animals. The PBS was then filtered sequentially through 5, 1, and 0.22 μm using a peristaltic pump. After that, 0.22 filter was treated using 1 mg/ml lysozyme and 0.2 mg/ml proteinase K (final concentrations) and total DNA was extracted with phenol/chloroform/isoamyl alcohol and chloroform/isoamyl alcohol, and DNA integrity was checked by agarose gel electrophoresis (Ghai et al., 2012).

ASSEMBLY AND GENERAL CHARACTERIZATION OF PHAGE GENOMES

We have sequenced by Illumina Hiseq 2000 using pair-end technology the DNA from the microbiota present in the epidermal mucosa of eels captured in Ebro Delta and Cabanes, and also the water sample, and by 454 FLX sequencer (Roche, Basel, Switzerland) the microbiota of mucus from eels, captured in Albufera de Valencia, farmed, and glass eels. These metagenomes have been deposited in NCBI SRA under the following accessions: SRR1578065, SRR1578068, SRR1578098, SRR1580820, SRR1580821, and SRR1580823. Metagenomic reads were assembled using Velvet (k–mer 51) and were annotated using Prodigal (Zerbino and Birney, 2008; Hyatt et al., 2010; Table S1). Only contigs bigger than 1 kb were considered for analysis. Annotation was refined manually using HHpred (Söding, 2005). The largest phage contigs were identified manually after annotation of contigs. The rest of them were fished using these contigs on a BLASTN search (Altschul et al., 1997). Moreover, CRISPR approach (see below) allowed us to find the prophage ProEnteroC171.

Phage genomes, classified as different families using ICTV, were downloaded from NCBI to accomplish a genomic comparison with our genomes. We used a previous protocol described (Mizuno et al., 2013). Comparisons were done using tBLASTX and BLOSUM45 matrix (Altschul et al., 1997). Minimum 30% sequence identity, 30 aa length and a maximum *e*-value of 10^{-3} were considered to filter the results. Phylogenetic tree was constructed using PHYLIP package (Felsenstein, 1993).

To find out if these contigs were closed, direct terminal repeats for phages were determined using Uniprot UGENE program (Okonechnikov et al., 2012). BLASTP search was used to identify the genes of interest to build the trees and the Ig-like domains. *Pseudomonas* phage phiKZ was used as the query in the ΦKZ genus specific searches. Finally, phylogenetic analyses using a group of genes (terminase large subunit, RNA and DNA polymerase [RNApol and DNApol], ribonucleoside diphosphate reductase alpha chain [*nrdA*], and phosphate starvation-induced protein [*phoH*]) allowed the taxonomic classification of these contigs. Trees were done using FastTree with 100 bootstraps (Price et al., 2009). All genes and genomes were download from Genbank, Pfam, or Uniprot database (Benson et al., 2005; Finn et al., 2006; Consortium, 2014).

HOST IDENTIFICATION

In order to identify the potential host for these phages CRISPR approach and tetranucleotide usage pattern (TUP) have been proved (Stern et al., 2012; Ogilvie et al., 2013). On one hand, three different approaches were used to search for putative CRISPR cassettes. (1) Protocol described by Stern et al. (2012) was tested, it consists in inferring the spacers using known CRISPR repeats and use these spacers to fish phage assembled contigs. To filter false positive spacers a minimum of 75 bp read length and 80% query coverage were used and to consider a spacer in a contig 100% identity and coverage were used. (2) Spacers were searched directly from assembled contigs using spacers from CRISPR database. Only matches over 80% coverage and 90% identity were considered. (3) CRISPRfinder tool was used to find CRISPR cassettes in assembled contigs.

One the other hand, spacers detected from different approaches were used to find viral contigs and putative hosts for those contigs. BLASTN search was done considering 100% identity as a filter. Finally, TETRA 1.0 was used to identify contigs with similar TUP (Teeling et al., 2004). Only contigs bigger than 10 kb and values of 0.6 or over were considered for analysis.

ABUNDANCE

In order to compare the abundance of viral and bacterial population in our datasets we count the number of reads recruited to viral and bacterial concatenated contigs. The number of reads was calculated using BLASTN and considering a minimum identity of 95 and a maximum *e*-value of 10^{-3} for filtering the results. The number of reads recruited against the genomes assembled was normalized per the size of the genome or concatenated (kb) and the dataset (Gb).

The abundance of those phages in other niches was analyzed comparing by TBLASTN the reads of marine (GOS, Albufera, Sargasso sea, Tampa bay, and Mediterranean bathypelagic habitat) and animal associated (mouse, termite, canine, cow, coral, and human) metagenomes against the viral proteins isolate in our data. Metagenomes were download from MG-RAST server (Meyer et al., 2008). A minimum identity of 60% and 10^{-5} *e*-value was considered for filtering the results. MG-RAST ID: 4440414.3, 4440440.3, 4440439.3, 4440413.3, 4440424.3, 4440422.3, 4440412.3, 4440411.3, 4440066.3, 4440062.3, 4440055.3, 4440056.3, 4440065.3, 4440063.3, 4440059.3, 4440064.3, 4483775.3, 4450680.3, 4450678.3, 4440284.3, 4440283.3, 4440285.3, 4440286.3, 4440102.3, 4444702.3, 4444703.3, 4444165.3, 4444164.3, 4440373.3, 4440375.3, 4440379.3, 4440377.3, 4440374.3, 4440381.3, 4440376.3, 4440378.3, 4440371.3, 4440370.3, 4440380.3, 4440372.3, 4447454.3, 4447455.3, 4447456.3, 4447457.3, 4447446.3, 4447447.3, 4447448.3, 4447449.3, 4441025.3, 4442464.3, 4441625.3, 4441625.4 4441627.3, 4441623.3, 4441624.3, 4441621.3, 4441622.3, 4441629.3, 4441628.3, 4441626.3, 4440330.3, 4440951.3, 4472804.3, 4472821.3, 4473347.3, 4473348.3, 4473365.3, 4473372.3, 4473378.3, 4473389.3, 4473411.3, 4473417.3, 4473438.3, and 4478542.3.

We also used VIROME database to search in the uploaded viromes for annotated ORFs to any of the viral genus found in our metagenomes (ΦKZ, ΦKMV, FelixO1like, and Φ16). The number

of ORFs was count using the online available tools in the VIROME website (Wommack et al., 2012).

PHAGE COUNTS

Samples from surrounding water and epidermal mucosa of eels farmed in tanks at 22°C and 1% salinity in facilities at University of Valencia (Planta de Acuarios de Experimentación, PAE) were collected. Samples were maintained on ice, sonicated in 3 pulses during 4 s. 50 and 3 ml of water and mucus, respectively, were directly filtered per 0.02 μm Anodisc polycarbonate filter (Whatman). Anodisc filters were stained with SYBR Green 5x, washed and visualized using epifluorescence microscope. For each sample, 25–30 images were analyzed using ImageJ (Schneider et al., 2012). Counts of bacteria and virus-like particles per milliliter were made using a previous protocol described (Patel et al., 2007).

RESULTS

We have previously analyzed the bacterial population present in the epidermal mucosa of farmed, wild, and glass eels by analysis of their metagenomes, which provided a first glimpse into the microbiota of the epidermal mucosa of eels (Carda-Diéguez et al., 2014). Interestingly, clear differences were observed in bacterial composition depending on the origin and the conditions where eels lived. Thus, *Pseudomonas* appeared as a common genus in eels and glass eels independently of their origin. In addition, *Vibrio*, *Shewanella*, *Aeromonas*, *Stenotrophomonas*, and *Acinetobacter* were the most abundant genera present in the microbiota from wild eels, while *Comamonas*, was only found in farmed eels (Carda-Diéguez et al., 2014).

Summary statistics for the sequenced metagenomes are provided in Table S1. The assembly of the Illumina datasets yielded a total of 17 viral contigs >10 kb and 23 contigs <10 kb, which were clearly related (>98% nucleotide identity) to the longer contigs (Figure S2). We selected five of these contigs for further analysis. From these metagenomes, we also assembled several contigs that appeared nearly identical to known bacterial genomes. Figure S3 shows an 893 kb contig that is >95% identical along its entire length to the genome of *Pseudomonas aeruginosa PA14* [originally isolated from a human patient (Lee et al., 2006)]. Long fragments of Caudovirales genomes have been detected in cellular metagenomes, probably due to the concatamer formation during lytic cycle, a natural process of genome amplification (Mizuno et al., 2013). In order to check if any of the assembled phages represented complete genomes, we performed all *vs* all comparisons of these viral contigs. We detected contigs overlapping in a circular fashion in the metagenome datasets from the 0.22 and 1 μm filters of the Ebro Delta (See Figure S2). This result suggests that all the genes in the genome had been captured and that these contigs represent complete phage genomes. Another method that may be useful in detecting the completeness of genomes is the presence of terminal repeats within the same contig. We identified a total of 4 complete phage genomes by these methods (See **Table 1** for details). Their analysis revealed that all correspond to tailed-bacteriophages. Additionally, nine putative prophages were also identified by the presence of host genes at one or both ends of the contig being their hosts *Stenotrophomonas*, *Achromobacter*, and *Enterobacteriaceae bacterium 9_2_54FAA* (Table S2 and Figure S4). Since we had multiple metagenomes, we checked for redundant contigs and found several examples (Figure S5). Even samples taken one year apart yielded identical phage contigs indicating a remarkable resilience and conservation.

MYOVIRUSES

Among tailed bacteriophages, Myoviruses are known because of their large sizes ranging from 11.6 to 358.6 kb. We found three Myoviruses that we named MyoΦKZC1, MyoΦKZC12, and MyoC35, with genome sizes of 220, 313, and 116 kb, respectively, (**Table 1**; **Figures 1** and **2**; Figure S6). MyoΦKZC12 and MyoΦKZC1 shared a protein with a putative CAAX protease domain and MyoC35 presented a putative ubiquitin-ligase, and the alpha and beta subunits of a proteasome complex. By complete genome analysis, MyoΦKZC12, and MyoΦKZC1 were predicted to be part of the ΦKZ genus (**Figure 3**). This genus is notorious for their genes involved in nucleotide metabolism (e.g., thymidylate synthase, thymidylate kinase, ribonucleoside diphosphate reductase subunit beta [*NrdB*] and alpha [*NrdA*], and dihydrofolate reductase, *RuvC* Holliday junction resolvasome; Cornelissen et al., 2012; Jang et al., 2013) all of which were found in MyoΦKZC12 and MyoΦKZC1 genomes. Phylogenetic analysis using DNApol (Figure S7), terminase large subunit (Figure S8) and ribonucleoside diphosphate reductase alpha chain (*nrdA*; Figure S9) genes also supported the taxonomic classification. The GC content of MyoΦKZC1 was in the expected range and MyoΦKZC12 reached the highest value ever published for this genus (58.22%; Table S3). Genomic comparison between *Pseudomonas* phage ΦKZ, PA7, and MyoΦKZC12 and MyoΦKZC1 revealed a low structural conservation together with a low nucleotide identity (Figure S6). In fact, only some structural protein, DNApol, RNApol, and terminase

Table 1 | Complete sequenced phages properties.

	Contig	Length (bp)	Genus	%GC	#ORF	%Annotated ORF
Cabanes	MyoΦKZC1	220.117	ΦKZ	45,19	227	32,15
Ebro Delta	MyoΦKZC12	313.980	ΦKZ	58,22	281	79,60
	MyoC35	116.788	FelixO1like	34,72	92	43,81
	PodoΦKMVC113	42.235	phiKMV-like (Podoviridae)	59,28	28	54,90
	MyoC197	40.198*	Unclassified Myoviridae	60,95	58	57

**incomplete genome.*

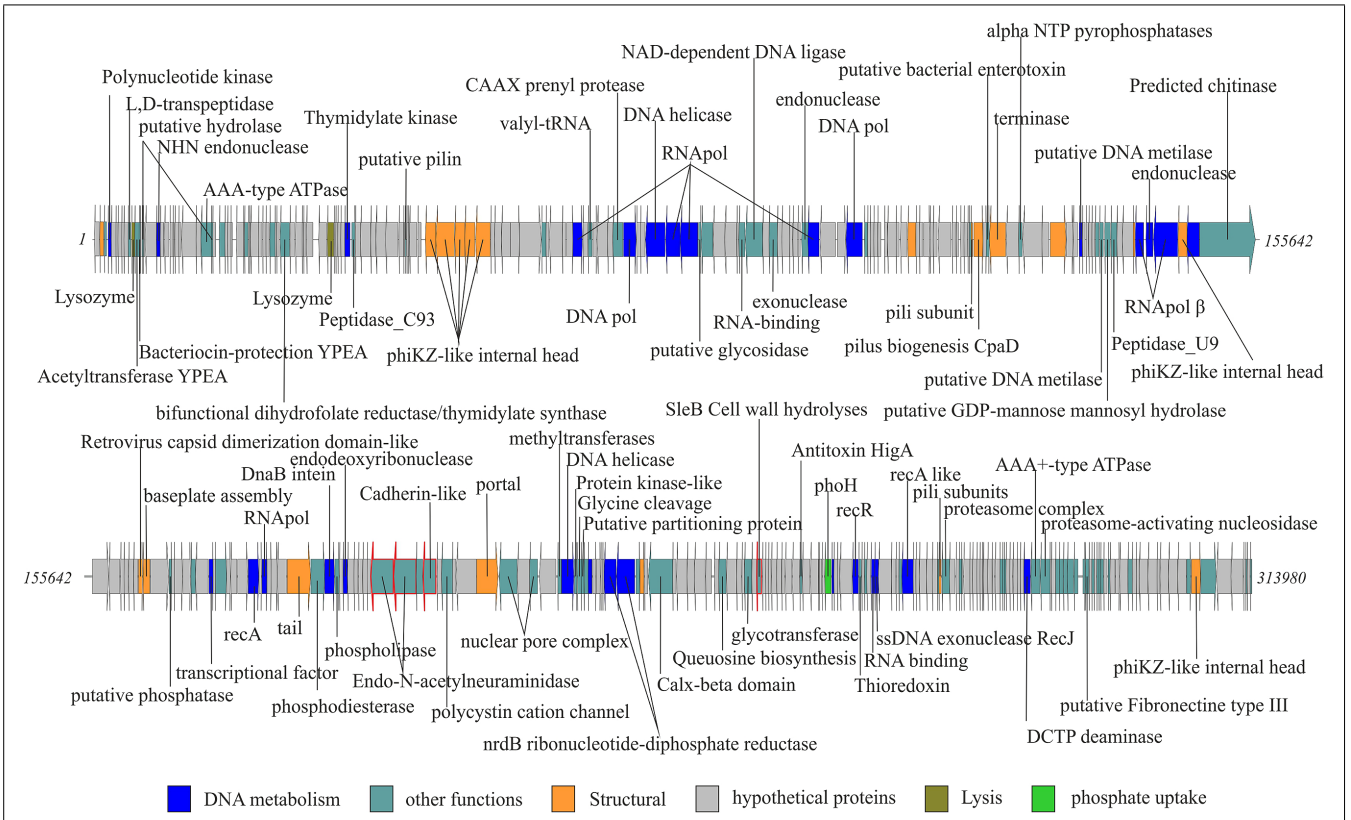


FIGURE 1 | MyoΦKZC12, complete new phage genome. Complete genome of the longest phage is represented in two rows. Gene functions ORFs are colored according to the color code. Ig-like domains are highlighted using a red fringe.

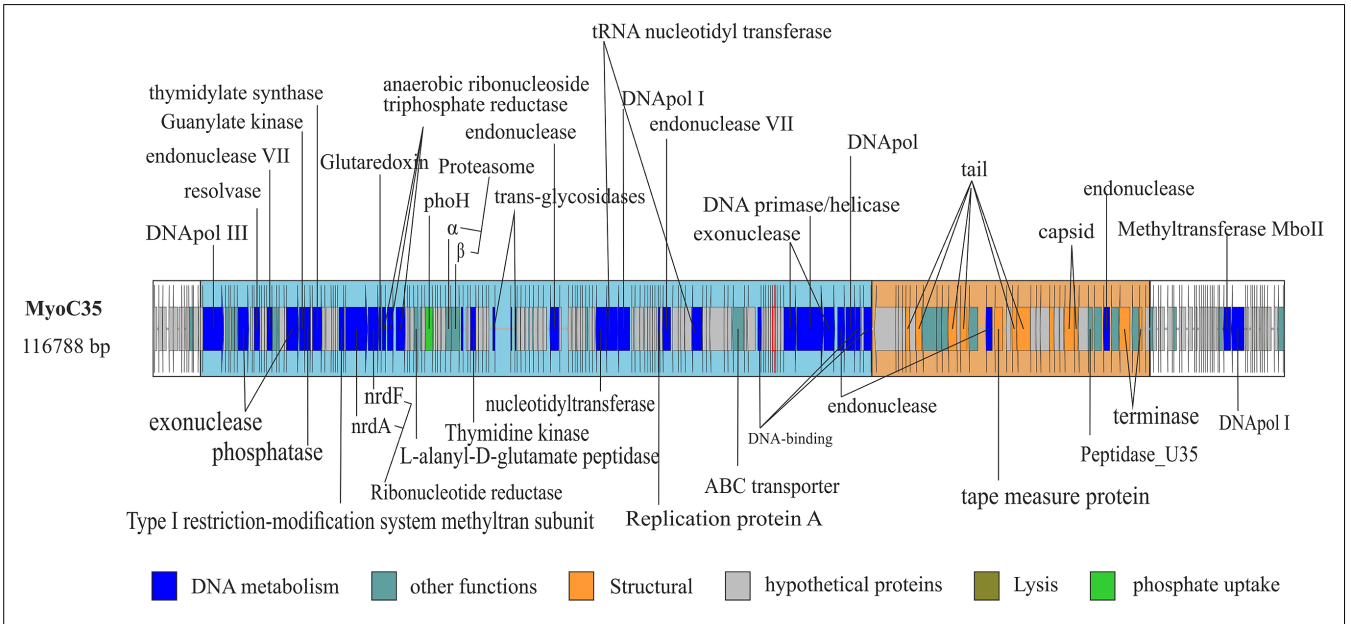


FIGURE 2 | Complete genome of MyoC35. Functional domains were highlighted using boxes colored according the function of the genes. White box indicate no functional domain. Ig-like domains are highlighted using a red fringe.

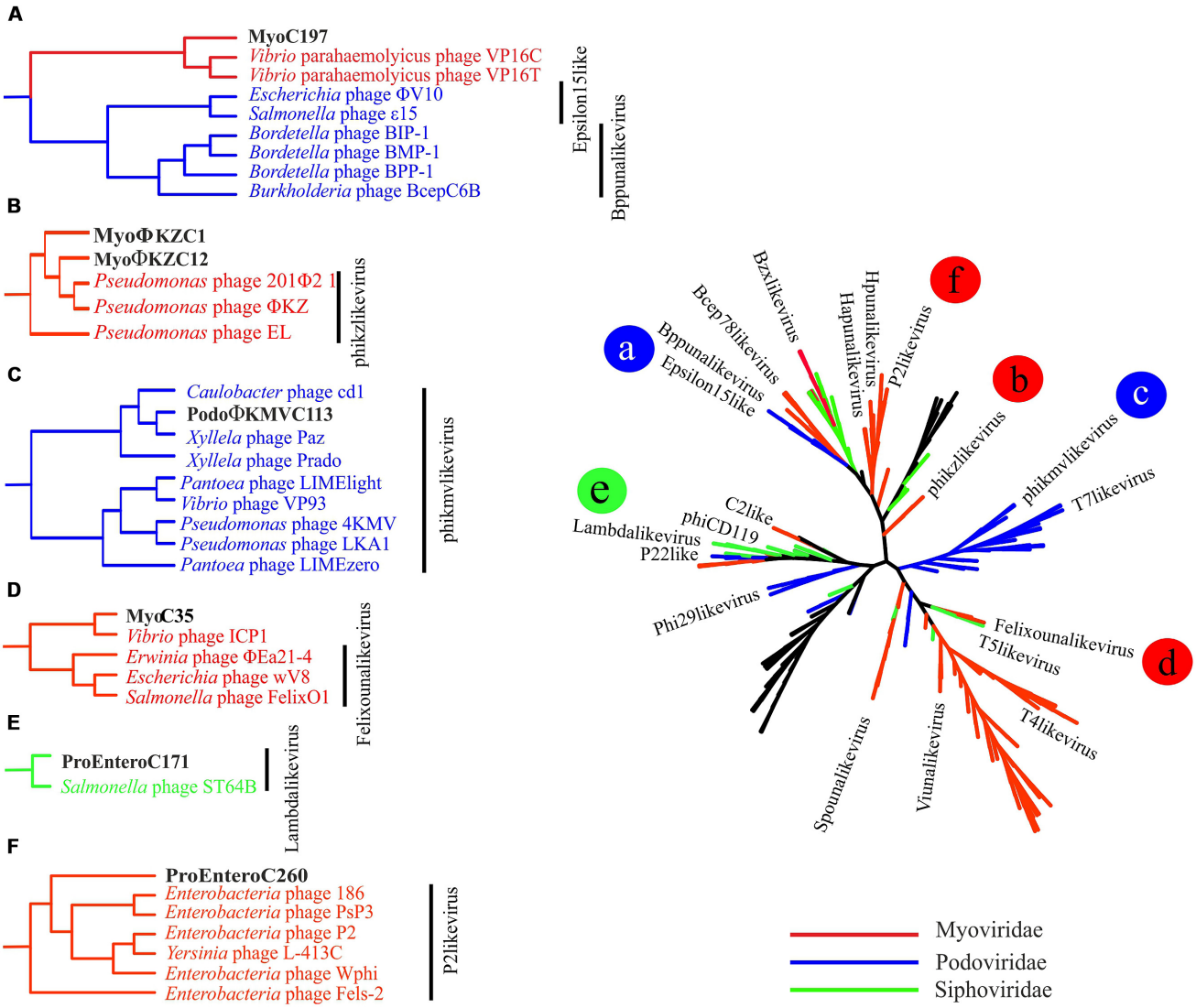


FIGURE 3 | Complete genomes phylogeny. All-vs-all comparison using NCBI phages genomes and the ones sequenced in our metagenome. The clustering was based on a sequence similarity derived metric. Colors were used to differentiate viral families. Branches in black indicate unclassified genomes. The branches in which the new complete genomes clustered are shown in detail on the left and marked with letters **(A–F)** that correspond with the position in the tree.

showed homology (30–60% aminoacid identity). Previous studies have shown a high rate of divergence among members of this genus (Cornelissen et al., 2012; Jang et al., 2013). Representatives of this genus target a variety of Gram-negative bacteria such as *Pseudomonas*, *Vibrio*, *Yersinia*, *Cronobacter*, *Salmonella*, and *Erwinia* (Table S3). We could not assign a putative host for these Myoviruses. Recent studies have shown that the eight RNA polymerase subunits from *Pseudomonas* phage ΦKZ form two polymerases: virion (vRNAP) and non-virion RNA polymerase (nvRNAP; Ceyssens et al., 2014). Moreover, Ceyssens et al. (2014) suggested that all subunits are present in all phages of the genus and the transcription is completely independent from the host. We have been able to identify some of them in MyoΦKZC12 and MyoΦKZC1 genomes. Further phylogenetic analysis using all RNApol subunits found in each ΦKZ genome showed a clustering of the eight different subunits in the tree (Figure S10).

MyoC35 (**Figure 2**, 116 kb) was classified as a FelixO1like phage by both whole genome comparison against NCBI (**Figure 3**) and clustering according to the terminase gene phylogenetic tree (Figure S8). In both analyses, MyoC35 clustered with *Cronobacter* phage vB_CsaM_GAP31 and *Vibrio* phage ICP1 although showed a low structural conservation and only DNApol, NrdA, and some structural proteins were similar (<50% identity; Figure S11). Those phages have not been assigned to any genus within the Myoviridae family. *Vibrio* phage ICP1 has been linked to FelixO1like genus by homology analysis using NrdA (Seed et al., 2011). Phylogenetic analysis using this gene clustered MyoC35

close to *Vibrio* phage ICP1 (Figure S9). However, the same analysis using DNApol clustered MyoC35 close to the Φ16 branch, a group recently added to NCBI which also appeared in this niche (see below; Figure S7). *Vibrio* phage ICP1 is a very specific *Vibrio* phage which has been isolated from stool samples from a cholera patient (Seed et al., 2011). In Ebro Delta metagenome, only small contigs were annotated as *V. cholerae*, however, *Vibrio* sp. *RC341* [closely related to *Vibrio cholerae* and *V. mimicus* (Haley et al., 2010)] was one of the most abundant organisms by number of contigs. They could be representatives of MyoC35 host.

Interestingly MyoΦKZC12 and MyoC35 genomes harbor a *phoH* gene, an auxiliary metabolic gene implicated in the regulation of phosphate uptake and metabolism under low-phosphate conditions (Hsieh and Wanner, 2010). The inclusion of this gene in the phylogenetic analysis supports the value of *phoH* as a signature gene for marine viruses (Figure S12; Goldsmith et al., 2011).

MyoC197 was predicted to be an incomplete myovirus with mixed characteristics (Figure S13, 40.2 kb). On one hand, this phage clustered with two *Vibrio* phages (VP16C and VP16T) by several phylogenetic analyses (**Figure 3**; Figures S7 and S8) although lacking the polypeptide deformylase gene and *vapE* (virulence associated protein), two typical genes of these *Vibrio* phages (Seguritan et al., 2003). On the other hand, MyoC197 contained several lambda phage related genes (genes for lambda head decoration protein D, major capsid protein E, and terminase large subunit GpA) and some Mu-like viruses related genes (genes for Mu-like prophage tail sheath protein, phage tail tube protein, FluMu protein Gp41, Mu-like prophage DNA circulation protein, and an uncharacterized protein conserved in bacteria DUF2313). In fact,

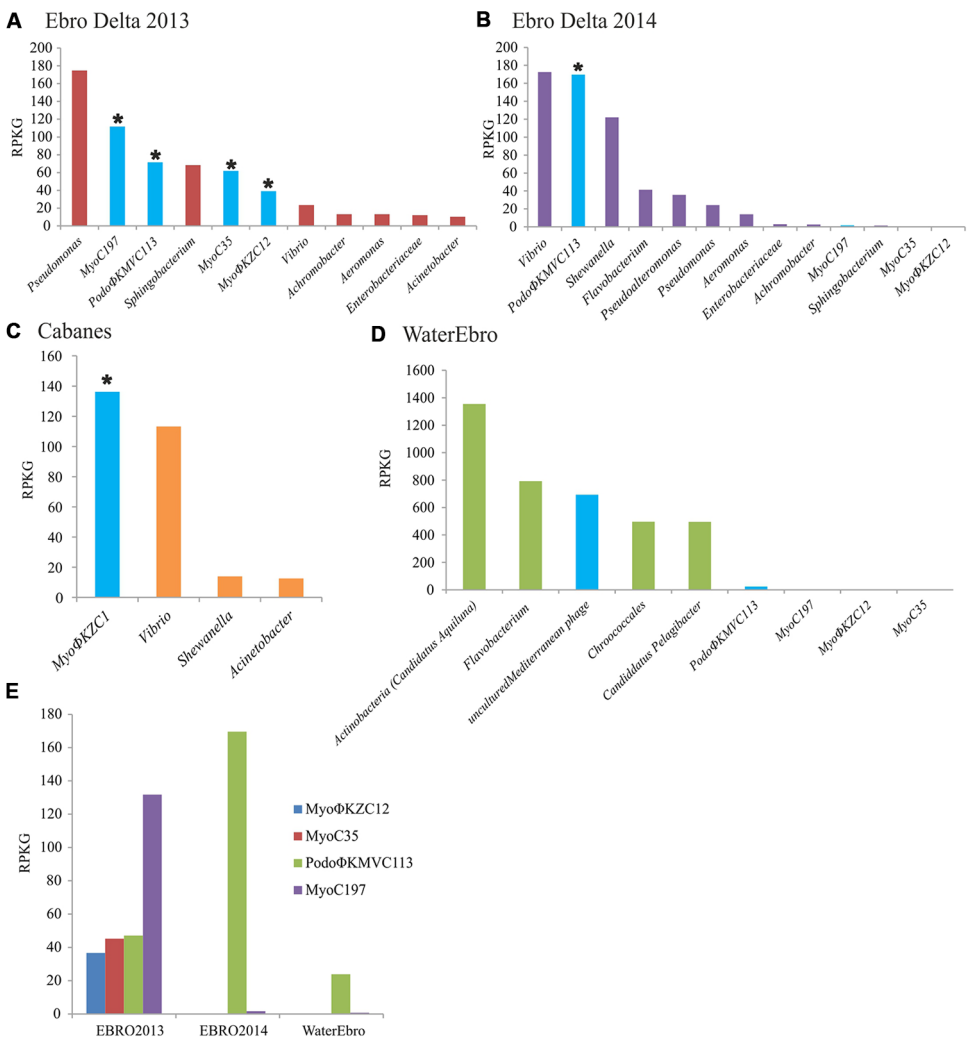


FIGURE 4 | Bacterial and viral abundance in the metagenomes. The abundance of bacterial genera and viral genomes were calculated by normalizing the number of reads recruited in their respective metagenomes per the size of the concatenated contigs (Kb) and the gigabases of the dataset (RPKG). Bacteriophage counts were highlighted in blue. Complete genomes were marked with an asterisk. **(A)** Ebro Delta 2013; **(B)** Ebro Delta 2014; **(C)** Cabanes-Torreblanca; **(D)** Water from Ebro Delta 2014; **(E)** Abundance of each phage genome in their respective metagenome; EBRO2013 and EBRO2014, metagenomes from Ebro Delta in 2013 and 2014, respectively; Water Ebro; metagenome from the water surrounding the eels from 2014.

25 of 58 genes in the contig had the highest similarity to *Vibrio* phage genes (50–70% identity; Figure S14). The *Vibrio* phages came from an environmental isolate of *V. parahaemolyticus* and are now classified as a new genus named Φ16 (Seguritan et al., 2003). Interestingly, MyoC197 was one of the most abundant viruses recruited from the Ebro Delta metagenome in 2013 (**Figure 4E**). Contigs annotated as *Vibrio* in metagenome from eels fished in Ebro Delta were assigned to *Vibrio sp. RC341*, *V. anguillarum*, *V. vulnificus,* and *V. fischeri* suggesting those species as putative hosts for MyoC197.

PODOVIRUSES

According to the phylogenetic analysis from the complete genome, PodoΦKMVC113 (Figure S13, 42.2 kb) is a Podovirus from the Autographivirinae subfamily and a putative new member of the ΦKMVlike genus of T7 related phages infecting *P. aeruginosa* (40.7–44.5 kb; **Figure 3**). RNApol, a hallmark gene for this genus, was found in PodoΦKMVC113, while single-strand interruptions (nicks) also typical in this genus were not present (Kulakov et al., 2009). Annotation results showed several genes assigned to other phages such as *Caulobacter* phage Cd1 and *Xanthomonas* phages. Phylogenetic analysis using the terminase and RNApol genes clustered PodoΦKMVC113 with *Caulobacter* phage Cd1, *Xylella* phages Paz, and Prado, all of them members of ΦKMVlike genus (Figures S8 and S10). As the other ΦKMV phages, PodoΦKMVC113 follows the structural organization based on three functional gene clusters encoded on the forward strand but with a rearrangement of lysis and structural domains (data not shown). A proteasome alpha subunit was also identified in this genome.

The segregation of the branch of PodoΦKMVC113, *Caulobacter* phage Cd1, and *Xylella* phages Prado and Paz suggests the formation of a new group very close to the ΦKMV genus (**Figure 3**). Interestingly, PodoΦKMVC113 was completely assembled in the same sampling point (Ebro Delta) a year later (Figure S4; identity 100%). Although no contig similar to PodoΦKMVC113 was assembled in the metagenome from the water, several reads were recruited against this genome (**Figure 4E**). It is probable that the complete genome would be assembled with larger sequencing efforts.

PHAGE ABUNDANCE IN MUCUS AND WATER AND PRESENCE OF SIMILAR PHAGES IN OTHER HABITATS

In order to assess the relative abundance of these phages in the different metagenomes, the contigs were compared with the raw reads. In all cases, the viral contigs recruited more than bacterial contigs except for *Pseudomonas* and *Sphingobium* in the sample from Ebro Delta (**Figures 4A,B**). In Cabanes, where the bacterial population was calculated to be composed by 60–80% of *Vibrio*, the number of reads recruited to MyoΦKZC1 genome was higher than those recruited to bacterial representatives (**Figure 4C**). Moreover, when these proportions were compared with the water metagenome, the relative abundance of phages and bacteria were reversed, suggesting that bacteriophages were retained or continuously produced in the epidermal mucosa (**Figure 4D**).

To test this apparently higher abundance of phages in mucus samples, we directly counted phages from mucus of farmed eels and the surrounding water by epifluorescence. As expected, the abundance of phages in mucus samples ($3.14{\cdot}10^6$ virus-like particles/ml) was higher than in the surrounding water ($1.47{\cdot}10^5$ virus-like particles/ml) confirming that mucus concentrates phages present in the water. We also counted bacteria and compared the ratio virus/bacteria in the mucus and water samples. Bacteria were more abundant in mucus ($1.62{\cdot}10^6$ vs. $1.49{\cdot}10^4$ bacteria-like particles per ml of mucus or water, respectively), suggesting that bacterial population was even more concentrated in the eel secretion. This led to a lower phage to bacteria ratio in the mucus (ca. 2:1) than in the water (ca. 10:1). Barr et al. (2013a) found 4.4 higher ratios of phages to bacteria in the mucus than in the surrounding environment (on average) in a study that included one teleost surface mucus. It is possible that our method to collect mucus by passive release from the fish (rather than by suction device as done in the mentioned paper) might have affected the results.

The relative abundance of phages in their respective samples was analyzed by recruitment as well. On one hand, MyoC197 was the genome that recruited the most while PodoΦKMVC-113, −C35, and −C12 reached similar but lower levels from Ebro Delta in 2013 (**Figure 4E**). On the other hand, a year later, PodoΦKMVC113 was the only one that increased its recruitment level while the Myoviruses almost disappeared. Considering that the bacterial population changed in this last sample, probably as a reflection of a change in water salinity (5.52–17.7 mS/cm; **Figures 4A,B**), the persistence of PodoΦKMVC113 and MyoC197 in the epidermal mucosa of eels in Ebro Delta for a year suggests that the host of these phages had remained in the epidermal mucosa while the host of the others Myoviruses did not. The bacterial genus that maintained and even raised its proportion in the epidermal mucosa through time was *Vibrio*. Therefore, this result suggests that *Vibrio* could be the host of PodoΦKMVC113 and MyoC197. Accordingly, the great reduction of the others Myoviruses mentioned could be related to a reduction in the host population. The decreasing numbers of reads recruited by *Pseudomonas, Achromobacter,* and *Sphingobium* make us consider them as putative hosts for these phages (**Figures 4A,B**). Finally, the only phage which recruited a significant number of reads in the metagenome from the water sample was PodoΦKMVC113.

Finally, metagenomes and metaviromes from marine and animal associated habitats were downloaded from MG-RAST in order to look for these phages in other niches. None of these datasets gave a number of matches high enough to consider that a similar virus was present in the searched metagenomes except for canine feces and cow rumen metagenomes in which reads similar to the ΦKZ and ΦKMV representatives were found. When reads identified as DNApol or RNApol were included to probe the presence of these genera in the rest of metagenomes, a single DNApol found in the cow rumen metagenome was similar to MyoΦKZC12 and MyoΦKZC1 and it clustered in the expected branch (Figure S7). Finally, two RNApol genes were recovered similar to MyoΦKZC12 from a canine metagenome and one to PodoΦKMVC113 from a cow metagenome that clustered in the respective branch (Figure S10). Moreover, we searched in the VIROME database for annotated open reading frames (ORFs) of

the different viral genus found in epidermal mucosa. Members of ΦKMV, FelixO1likevirus, and Φ16 were found in practically all marine, soil, and host-associated viromes while ΦKZ genus was present in very low amounts in some marine viromes but was highly represented in host-associated ones, especially in the cow rumen virome. ΦKZ members have been isolated from bacterial pure cultures from a great diversity of environments: sewage, pond water, compost, soil, chicken feces, fresh, and marine water, but this is the first time that they have been sequenced from an environmental habitat. It is noteworthy that 40 contigs from Albufera metagenome were annotated as ΦKZ (30–60% identity). Furthermore, DNApol and RNApol subunits were found in some of these contigs and included in the analysis confirming the annotation (Figures S7 and S10). This turned ΦKZ in the only genus found in all three metagenomes from mucus of wild eels.

CONCLUSION, VIRAL POPULATION, AND FUNCTIONAL ROLE

Most of the metagenomic studies have been carried out with human or water related environments (Venter et al., 2004; Daniel, 2005; Ghai et al., 2011, 2012; Yu and Zhang, 2012; Lipson et al., 2013; Mizuno et al., 2013). There are only few studies of the microbiota associated to animals and even less to wild animals (Vega Thurber et al., 2009; Swanson et al., 2011; Marcobal and Sonnenburg, 2012; Parfrey and Knight, 2012; Ross et al., 2012, 2013; Schwartz et al., 2012; Singh et al., 2012; Wong and Rawls, 2012; Bodewes et al., 2014). In this study, we analyzed the viral content of the epidermal mucosa of the eel to test the hypothesis that it concentrates phages present in water, which in turn, control bacterial populations, including those of pathogenic species. We found a replicating viral community in this niche formed by Myovirus and Podovirus. In addition, we have found evidences that ΦKZ genus and the Podovirus could be part of the resident microbiota associated to the eel mucosal surface.

The recovery of large contigs of genomes from metagenomic samples is normally an indication of lower diversity. In this sense, the assembly of large genomic tracts from this animal mucus samples indicates low diversity what in itself is remarkable since this external layer is immersed in waters with rich microbial communities. If it actually is a defense mechanism for the protection of the fish it remains to be demonstrated. On the other hand, we found evidence for large amounts of phages. There is little doubt that the genomes retrieved derive from actively replicating phages in cells present in the samples. In addition, despite the abundance of potential pathogens found in the metagenomes (especially, in wild eels: *Pseudomonas aeruginosa*, *Aeromonas veronii, V. anguillarum, Acinetobacter baumanii, Achromobacter xylosoxydans,* etc) we did not observe clinical symptoms of infectious disease in the captured eels.

All these findings support the BAM model described by Barr et al. (2013a) that proposes that phages attached to the mucous are ideally located to infect bacterial cells attracted by the rich nutrients provided by the animal. Although epifluorescent counts gave in our case lower phage/bacteria ratios in the mucus that in the water, we could detect a clear enrichment of phages in the mucus (20 times more). The passive way to obtain mucus that does not probably retrieves the mucus layers most closely associated to the fish surface might explain this discrepancy.

Phages use Ig-like domains present in the capsids to attach to mucin. However, only a single putative Ig-domain was found in MyoΦKZC12 and a fibronectin type III domain in MyoΦKZC1 phage. Considering that the Barr study was performed in T4 group bacteriophages and MyoΦKZC12 and MyoΦKZC1 are members of a genus classified as T4-like, it is possible that attachment through Ig-like domains is specific of T4 bacteriophages and the rest use a different protein. When others Ig-like domains and domains that are known to bind to carbohydrates, which could mediate BAM-mechanisms were searched in the sequenced phages, a C-type lectin domain was found in MyoΦKZC1 and MyoC35. Moreover, three and one Ig-like domains annotated in Pfam as invasion/intimin cell adhesion were found in MyoΦKZC12 and MyoC197, respectively. Finally, we found BACON domain (Bacteroidetes-Associated Carbohydrate-binding Often N-terminal) in a tail protein of MyoC197.

Recent work related to bacterial infection in aquaculture proposes phage therapy as a preventive strategy (Vinod et al., 2006; Mateus et al., 2014; Rong et al., 2014). Along these lines, phage therapy using mucus associated phage communities against potential pathogens could be a useful concept in European eel farming.

ACKNOWLEDGMENTS

This work has been financed by grants AGL2011-29639 (cofunded with FEDER funds) and Programa Consolider-Ingenio 2010 CSD2009-00006 from MICINN. We thank the Servicio de Vida Silvestre (DG de Medio Natural, Conselleria de Infrestructuras, Territorio y Medio Ambiente) for helping in sampling of wild eels. Miguel Carda-Diéguez thanks MICINN for the Fellowship FPI (BES-2012-052361).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00003/abstract

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST‘ new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., et al. (2013a). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Barr, J. J., Youle, M., and Rohwer, F. (2013b). Innate and acquired bacteriophage-mediated immunity. *Bacteriophage* 3, e25857. doi: 10.4161/bact.25857

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). GenBank. *Nucleic Acids Res.* 33, D34–D48. doi: 10.1093/nar/gki063

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C. M. E., van den Brand, J. M. A., Osterhaus, A. D. M. E., and Smits, S. L. (2014). Viral metagenomic analysis of feces of wild small carnivores. *Virol. J.* 11, 89. doi: 10.1186/1743-422X-11-89

Carda-Diéguez, M., Ghai, R., Rodriguez-Valera, F., and Amaro, C. (2014). Metagenomics of the mucosal microbiota of european eels. *Genome Announc.* 2: e01132-14. doi: 10.1128/genomeA.01132-14

Ceyssens, P.-J., Minakhin, L., Van den Bossche, A., Yakunina, M., Klimuk, E., Blasdel, B., et al. (2014). Development of giant bacteriophage ΦKZ is independent of the host transcription apparatus. *J. Virol.* 88, 10501–10510. doi: 10.1128/JVI.01347-14

Consortium, T. U. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42, D191–D198. doi: 10.1093/nar/gkt1140

Cornelissen, A., Hardies, S. C., Shaburova, O. V., Krylov, V. N., Mattheus, W., Kropinski, A. M., et al. (2012). Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse ΦKZ-related phages. *J. Virol.* 86, 1844–1852. doi: 10.1128/JVI.06330-11

Daniel, R. (2005). The metagenomics of soil. *Nat. Rev. Microbiol.* 3, 470–478. doi: 10.1038/nrmicro1160

Ellis, A. (2001). Innate host defense mechanisms of fish against viruses and bacteria. *Dev. Comp. Immunol.* 25, 827–839. doi: 10.1016/S0145-305X(01)00038-6

Felsenstein, J. (1993). *PHYLIP Phylogeny Inference Package*, version 3.5 c.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res.* 34, D247–D251. doi: 10.1093/nar/gkj149

Ghai, R., Hernandez, C. M., Picazo, A., Mizuno, C. M., Ininbergs, K., Díez, B., et al. (2012). Metagenomes of Mediterranean coastal lagoons. *Sci. Rep.* 2, 490. doi: 10.1038/srep00490

Ghai, R., Rodriguez-Valera, F., McMahon, K. D., Toyama, D., Rinke, R., Cristina Souza de Oliveira, T., et al. (2011). Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS ONE* 6:e23785. doi: 10.1371/journal.pone.0023785

Goldsmith, D. B., Crosti, G., Dwivedi, B., McDaniel, L. D., Varsani, A., Suttle, C. A., et al. (2011). Development of phoH as a novel signature gene for assessing marine phage diversity. *Appl. Environ. Microbiol.* 77, 7730–7739. doi: 10.1128/AEM.05531-11

Haley, B. J., Grim, C. J., Hasan, N. A., Choi, S.-Y., Chun, J., Brettin, T. S., et al. (2010). Comparative genomic analysis reveals evidence of two novel *Vibrio species* closely related to V. *cholerae*. *BMC Microbiol.* 10:154. doi: 10.1186/1471-2180-10-154

Hsieh, Y.-J., and Wanner, B. L. (2010). Global regulation by the seven-component Pi signaling system. *Curr. Opin. Microbiol.* 13, 198–203. doi: 10.1016/j.mib.2010.01.014

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. doi: 10.1371/journal.pone.0057355

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 Bioinform.* 11:119. doi: 10.1186/1471-2105-11-119

Jang, H. B., Fagutao, F. F., Nho, S. W., Park, S. B., Cha, I. S., Yu, J. E., et al. (2013). Phylogenomic network and comparative genomics reveal a diverged member of the ΦKZ-related group, marine vibrio phage ΦJM-2012. *J. Virol.* 87, 12866–12878. doi: 10.1128/JVI.02656-13

Kulakov, L. A., Ksenzenko, V. N., Shlyapnikov, M. G., Kochetkov, V. V., Del Casale, A., Allen, C. et al. (2009). Genomes of "phiKMV-like viruses" of *Pseudomonas aeruginosa* contain localized single-strand interruptions. *Virology* 391, 1–4. doi: 10.1016/j.virol.2009.06.024

Lee, D. G., Urbach, J. M., Wu, G., Liberati, N. T., Feinbaum, R. L., Miyata, S., et al. (2006). Genomic analysis reveals that *Pseudomonas aeruginosa* virulence is combinatorial. *Genome Biol.* 7, R90. doi: 10.1186/gb-2006-7-10-r90

Lee, W.-C., Chen, Y.-H., Lee, Y.-C., and Liao, I. C. (2003). The competitiveness of the eel aquaculture in Taiwan, Japan, and China. *Aquaculture* 221, 115–124. doi: 10.1016/S0044-8486(03)00004-8

Lipson, D. A., Haggerty, J. M., Srinivas, A., Raab, T. K., Sathe, S., and Dinsdale, E. A. (2013). Metagenomic insights into anaerobic metabolism along an Arctic peat soil profile. *PLoS ONE* 8:e64659. doi: 10.1371/journal.pone.0064659

Marcobal, A., and Sonnenburg, J. L. (2012). Human milk oligosaccharide consumption by intestinal microbiota. *Clin. Microbiol. Infect.* 18(Suppl. 4), 12–15. doi: 10.1111/j.1469-0691.2012.03863.x

Mateus, L., Costa, L., Silva, Y. J., Pereira, C., Cunha, A., and Almeida, A. (2014). Efficiency of phage cocktails in the inactivation of Vibrio in aquaculture. *Aquaculture* 424–425, 167–173. doi: 10.1016/j.aquaculture.2014.01.001

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9:386. doi: 10.1186/1471-2105-9-386

Minot, S., Sinha, R., Chen, J., and Li, H. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* 21, 1616–1625. doi: 10.1101/gr.122705.111.1616

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLoS Genet.* 9:e1003987. doi: 10.1371/journal.pgen.1003987

Nguyen-Kim, H., Bouvier, T., Bouvier, C., Doan-Nhu, H., Nguyen-Ngoc, L., Rochelle-Newall, E., et al. (2014). High occurrence of viruses in the mucus layer of scleractinian corals. *Environ. Microbiol. Rep.* 6, 675–682. doi: 10.1111/1758-2229.12185

Ogilvie, L. A., Bowler, L. D., Caplin, J., Dedi, C., Diston, D., Cheek, E., et al. (2013). Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nat. Commun.* 4, 2420. doi: 10.1038/ncomms3420

Okonechnikov, K., Golosova, O., and Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi: 10.1093/bioinformatics/bts091

Parfrey, L. W., and Knight, R. (2012). Spatial and temporal variability of the human microbiota. *Clin. Microbiol. Infect.* 18(Suppl. 4), 8–11. doi: 10.1111/j.1469-0691.2012.03861.x

Patel, A., Noble, R. T., Steele, J. A., Schwalbach, M. S., Hewson, I., and Fuhrman, J. A. (2007). Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nat. Protoc.* 2, 269–276. doi: 10.1038/nprot.2007.6

Phan, T. G., Kapusinszky, B., Wang, C., Rose, R. K., Lipton, H. L., and Delwart, E. L. (2011). The fecal viral flora of wild rodents. *PLoS Pathog.* 7:e1002218. doi: 10.1371/journal.ppat.1002218

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Rodriguez-Valera, F., Mizuno, C. M., and Ghai, R. (2014). Tales from a thousand and one phages. *Bacteriophage* 4, e28265. doi: 10.4161/bact.28265

Rong, R., Lin, H., Wang, J., Khan, M. N., and Li, M. (2014). Reductions of *Vibrio parahaemolyticus* in oysters after bacteriophage application during depuration. *Aquaculture* 418–419, 171–176. doi: 10.1016/j.aquaculture.2013.09.028

Ross, E. M., Moate, P. J., Bath, C. R., Davidson, S. E., Sawbridge, T. I., Guthridge, K. M., et al. (2012). High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. *BMC Genet.* 13:53. doi: 10.1186/1471-2156-13-53

Ross, E. M., Petrovski, S., Moate, P. J., and Hayes, B. J. (2013). Metagenomics of rumen bacteriophage from thirteen lactating dairy cattle. *BMC Microbiol.* 13:242. doi: 10.1186/1471-2180-13-242

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9, 671–675. doi: 10.1038/nmeth.2089

Schwartz, S., Friedberg, I., Ivanov, I. V., Davidson, L. A., Goldsby, J. S., Dahl, D. B., et al. (2012). A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol.* 13, r32. doi: 10.1186/gb-2012-13-4-r32

Seed, K., Bodi, K., and Kropinski, A. (2011). Evidence of a dominant lineage of *Vibrio cholerae*-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. *MBio* 2, 1–9. doi: 10.1128/mBio.00334-10

Seguritan, V., Feng, I., and Rohwer, F. (2003). Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. *J. Bacteriol.* 185, 6434–6447. doi: 10.1128/JB.185.21.6434

Singh, K. M., Jakhesara, S. J., Koringa, P. G., Rank, D. N., and Joshi, C. G. (2012). Metagenomic analysis of virulence-associated and antibiotic resistance genes of microbes in rumen of Indian buffalo (Bubalus bubalis). *Gene* 507, 146–151. doi: 10.1016/j.gene.2012.07.037

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951–960. doi: 10.1093/bioinformatics/bti125

Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* 22, 1985–1994. doi: 10.1101/gr.138297.112

Swanson, K. S., Dowd, S. E., Suchodolski, J. S., Middelbos, I. S., Vester, B. M., Barry, K. A., et al. (2011). Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *ISME J.* 5, 639–649. doi: 10.1038/ismej.2010.162

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. O. (2004). TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinform.* 5:163. doi: 10.1186/1471-2105-5-163

Tesch, F.-W. (2003). "Body structure and functions," in *The Eel,* 5th Edn, ed. J. E. Thorpe (Oxford: Blackwell Science).

Valiente, E., Lee, C. T., Lamas, J., Hor, L., and Amaro, C. (2008). Role of the virulence plasmid pR99 and the metalloprotease Vvp in resistance of *Vibrio vulnificus* serovar E to eel innate immunity. *Fish Shellfish Immunol.* 24, 134–141. doi: 10.1016/j.fsi.2007.10.007

Vega Thurber, R., Willner-Hall, D., Rodriguez-Mueller, B., Desnues, C., Edwards, R. A., Angly, F., et al. (2009). Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* 11, 2148–2163. doi: 10.1111/j.1462-2920.2009.01935.x

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun Sequencing of the sargasso sea. *Science* 304, 66–74.

Vinod, M. G., Shivu, M. M., Umesha, K. R., Rajeeva, B. C., Krohne, G., Karunasagar, I., et al. (2006). Isolation of *Vibrio harveyi* bacteriophage with a potential for biocontrol of luminous vibriosis in hatchery environments. *Aquaculture* 255, 117–124. doi: 10.1016/j.aquaculture.2005.12.003

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

Wong, S., and Rawls, J. F. (2012). Intestinal microbiota composition in fishes is influenced by host ecology and environment. *Mol. Ecol.* 21, 3100–3102. doi: 10.1111/j.1365-294X.2012.05646.x

Yu, K., and Zhang, T. (2012). Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS ONE* 7:e38183. doi: 10.1371/journal.pone.0038183

Zablocki, O., van Zyl, L., Adriaenssens, E. M., Rubagotti, E., Tuffin, M., Cary, C., et al. (2014). High diversity of tailed phages, eukaryotic viruses and virophage-like elements in the metaviromes of Antarctic soils. *Appl. Environ. Microbiol.* 81, 1–29. doi: 10.1128/AEM.01525-14

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

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

Received: 05 December 2014; accepted: 03 January 2015; published online: 29 January 2015.

Citation: Carda-Diéguez M, Mizuno CM, Ghai R, Rodriguez-Valera F and Amaro C (2015) Replicating phages in the epidermal mucosa of the eel (Anguilla anguilla). Front. Microbiol. ***6****:3. doi: 10.3389/fmicb.2015.00003*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 March 2015
doi: 10.3389/fmicb.2015.00185

RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi

James Angus Chandler[*†], *Rachel M. Liu*[†] and *Shannon N. Bennett*[*]

Department of Microbiology, California Academy of Sciences, San Francisco, CA, USA

Mosquitoes, most often recognized for the microbial agents of disease they may carry, harbor diverse microbial communities that include viruses, bacteria, and fungi, collectively called the microbiota. The composition of the microbiota can directly and indirectly affect disease transmission through microbial interactions that could be revealed by its characterization in natural populations of mosquitoes. Furthermore, the use of shotgun metagenomic sequencing (SMS) approaches could allow the discovery of unknown members of the microbiota. In this study, we use RNA SMS to characterize the microbiota of seven individual mosquitoes (species include *Culex pipiens, Culiseta incidens,* and *Ochlerotatus sierrensis*) collected from a variety of habitats in California, USA. Sequencing was performed on the Illumina HiSeq platform and the resulting sequences were quality-checked and assembled into contigs using the A5 pipeline. Sequences related to single stranded RNA viruses of the *Bunyaviridae* and *Rhabdoviridae* were uncovered, along with an unclassified genus of double-stranded RNA viruses. Phylogenetic analysis finds that in all three cases, the closest relatives of the identified viral sequences are other mosquito-associated viruses, suggesting widespread host-group specificity among disparate viral taxa. Interestingly, we identified a *Narnavirus* of fungi, also reported elsewhere in mosquitoes, that potentially demonstrates a nested host-parasite association between virus, fungi, and mosquito. Sequences related to 8 bacterial families and 13 fungal families were found across the seven samples. *Bacillus* and *Escherichia/Shigella* were identified in all samples and *Wolbachia* was identified in all *Cx. pipiens* samples, while no single fungal genus was found in more than two samples. This study exemplifies the utility of RNA SMS in the characterization of the natural microbiota of mosquitoes and, in particular, the value of identifying all microbes associated with a specific host.

Keywords: metagenomics, shotgun sequencing, microbiota, *Culex pipiens, Culiseta, Ochlerotatus*, *Bunyaviridae*, *Rhabdoviridae*

OPEN ACCESS

Edited by:
Katrine L. Whiteson, University of California, Irvine, USA

Reviewed by:
Thawornchai Limjindaporn, Mahidol University, Thailand
Lark L. Coffey, University of California, Davis, USA

***Correspondence:**
James Angus Chandler and Shannon N. Bennett, Department of Microbiology, California Academy of Sciences, 55 Music Concourse Drive, Golden Gate Park, San Francisco, CA 94118, USA
jchandler@calacademy.org; sbennett@calacademy.org

[†]*These authors have contributed equally to this work.*

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *03 December 2014*
Accepted: *19 February 2015*
Published: *24 March 2015*

Citation:
Chandler JA, Liu RM and Bennett SN (2015) RNA shotgun metagenomic sequencing of northern California (USA) mosquitoes uncovers viruses, bacteria, and fungi. Front. Microbiol. 6:185. doi: 10.3389/fmicb.2015.00185

Introduction

The diversity of animal-associated microbes, collectively called the microbiota, and their ubiquitous role in host ecology, physiology, and evolution has reached new levels of appreciation with the advent of next-generation sequencing technologies (Foster et al., 2012; McFall-Ngai et al., 2013).

Culture-dependent methods are limited in their capacity to capture a full picture of the microorganisms present, as only a minority of microbes are easily culturable under "standard" laboratory conditions (Amann et al., 1995 but see Donachie et al., 2007). Metagenomic PCR circumvents this issue by identifying organisms directly from environmentally acquired nucleic acids, often using taxon-specific primers that target genes such as the ribosomal small subunit (Schmidt et al., 1991; Handelsman et al., 1998). However, this method has known issues related to primer selection, chimera formation, and gene copy number (Ashelford et al., 2006; Kembel et al., 2012; Ghyselinck et al., 2013; Klindworth et al., 2013). Shotgun metagenomic sequencing (SMS), which does not rely upon an initial PCR step, can avoid the limitations of culture-dependent and PCR-based methods (Venter et al., 2004; Haas et al., 2011). Combined with high throughput sequencing technologies, SMS has been successfully used to characterize the microbiota of honeybees (Runckel et al., 2011; Engel et al., 2012), termites (Warnecke et al., 2007), humans (Qin et al., 2010), and a variety of mammals (Muegge et al., 2011). SMS has an added benefit that the data produced is not restricted to a single taxon (i.e., only bacteria or only fungi) and has been used to identify a variety of microorganisms associated with a single host (Runckel et al., 2011). In this study, we use SMS and high throughput sequencing to examine the microbiota of mosquitoes collected in northern California, USA.

Mosquitoes are vectors of many clinically and economically important diseases, such as malaria and dengue (Morens et al., 2004; Fukuda et al., 2011). Additionally, mosquitoes carry many non-pathogenic microbes (Minard et al., 2013) and some of these are prime candidates for symbiont mediated transmission disruption. For example, the intracellular bacterium *Wolbachia* can affect dengue transmission in *Aedes aegypti* (Walker et al., 2011) and reduces the titer of *West Nile virus* in *Culex quinquefasciatus* (Glaser and Meola, 2010). Genetically modifying bacteria that are naturally associated with mosquitoes is also a promising approach to transmission disruption, as has been shown with *Anopheles* mosquitoes and malaria (Wang et al., 2012). These studies and others (reviewed in Cirimotich et al., 2011) were made possible through a previous understanding of the mosquito microbiota.

Numerous culture-dependent and culture-independent studies have examined the bacterial communities associated with mosquitoes (reviewed in Minard et al., 2013). Generally, the bacterial taxa Gammaproteobacteria and Firmicutes are major components of the mosquito microbiota, but other groups, such as the Alphaproteobacteria and the Betaproteobacteria, are also present (Minard et al., 2013). Studies investigating the mosquito microbiota have shown that its composition is based on both environmental (such as habitat) and host-intrinsic (such as sex, species, and developmental stage) factors (Minard et al., 2013). The function of the microbiota remains unclear in many cases, although its experimental removal does arrest larval development in three different mosquito species (Coon et al., 2014).

As with other animal hosts [e.g., humans (Huffnagle and Noverr, 2013) and *Drosophila* (Broderick and Lemaitre, 2012)], fungi associated with mosquitoes are relatively understudied compared to their bacterial counterparts. Most previous work investigating the fungi associated with mosquitoes has focused on entomopathogenic fungi and their use in mosquito control (Scholte et al., 2004; de Faria and Wraight, 2007). However, to our knowledge, there are no studies that have used culture-independent techniques to characterize the total fungal communities of natural populations of mosquitoes.

Shotgun metagenomic sequencing has previously been used to investigate the mosquito microbiota. In particular, the viral diversity associated with wild-caught mosquitoes has been characterized and these studies have shown that SMS is sufficient to recover and identify viruses from a variety of taxonomic groups (Ma et al., 2011; Ng et al., 2011; Cook et al., 2013; Chandler et al., 2014). Furthermore, mosquitoes that have been laboratory infected with known mosquito-vectored viruses (such as dengue, yellow fever, or chikungunya) have been subjected to SMS (Bishop-Lilly et al., 2010; Hall-Mendelin et al., 2013). In addition to successful identification of the viruses and mosquitoes, genetic material from both bacteria and fungi were recovered via SMS.

Lab-infected mosquito studies (Bishop-Lilly et al., 2010; Hall-Mendelin et al., 2013) suggest the feasibility of using SMS to identify both viruses and other microorganisms (such as bacteria and fungi) in wild populations of mosquitoes, however, to our knowledge, this has never been done. Because symbiont mediated transmission disruption is an emerging tool in controlling vector-borne diseases (reviewed in Cirimotich et al., 2011; Weiss and Aksoy, 2011), SMS of wild mosquitoes may be particularly informative because it simultaneously identifies both viruses and other symbionts and could uncover any correlation between the two in the same host. This study represents a proof of concept in characterizing microbiota using a taxonomically broad approach that could ultimately prove useful in exposing significant novel interactions between microbes. In particular, we use SMS to characterize the microbiota of seven individual mosquitoes of three different species that were collected from a variety of natural habitats in northern California, USA, which is an area where *West Nile virus* has been reported. Using RNA-based SMS on the Illumina platform, we identify sequences related to viruses, bacteria, and fungi in each individual. Furthermore, we were able to verify mosquito species identities using SMS data alone. This work exemplifies the utility of SMS to study the natural microbiota of mosquitoes and we hope it prompts future research in this area.

Materials and Methods

Collection and Identification

Several hundred mosquitoes were collected from seven locations in northern California (Pepperwood Preserve, Bolinas, Stinson Beach, San Rafael, Mill Valley, San Francisco, and San Mateo) during March–November 2013 (**Figure 1**). Locations were chosen to represent a range of habitats, from sylvatic/wild (e.g., Pepperwood Preserve) to urban (e.g., San Francisco), (**Table 1**). Collections occurred under the permit and permission

FIGURE 1 | Collection locations and a summary of the microbial sequences identified in each sample. Locations A through G correspond to the samples as listed in **Table 1**. A summary of the viral, bacterial, and fungal sequences identified in each sample is included. Details for each microbial group can be found in **Tables 2–4**. The map was created using Google Earth Pro, with data from the U.S. Navy, LDEO-Columbia, NSF, NOAA, SIO, NGA, MBARI, and GEBCO.

agreements of the Marin/Sonoma Mosquito and Vector Control District, or on private lands with the owners' permission.

We employed several methods to collect mosquitoes, including hand nets, gravid traps baited with hay-infused water, or Zumba^{TM} traps baited with carbon dioxide gas (CO_2), heat packs, and SkinLure^{TM}. All samples were frozen within 24–48 h of being trapped and stored at −80°C. We identified mosquitoes morphologically using a dissecting microscope and key (Bohart and Washino, 1978), retaining voucher specimens for each species and trapping event. In several cases, morphological

TABLE 1 | Sample details.

Map	Location	Habitat	Species	Library name	Total reads	Number of reads after quality control	Number of contigs
A	Pepperwood Preserve	Sylvatic/wild	*Ochlerotatus sierrensis*	pepp.ochl	35,807,449	32,066,403	29,911
B	Bolinas	Rural	*Culex pipiens*	boli.cpip	32,644,864	29,446,138	30,686
C	Stinson Beach	Rural	*Culex pipiens*	stin.cpip	31,123,780	27,519,860	29,234
D	San Rafael	Urban	*Culex pipiens*	sraf.cpip	36,917,070	32,872,651	44,558
E	Mill Valley	Rural/suburban	*Culiseta incidens*	mill.culi	36,248,881	32,552,420	13,577
F	San Francisco	Urban	*Culex pipiens*	sfra.cpip	33,906,608	30,669,364	53,542
G	San Mateo	Suburban/urban	*Culiseta incidens*	smat.culi	35,641,996	32,203,598	110,242

TABLE 2 | Viral sequences identified in northern California mosquitoes.

Location	Species	Contig designation	Length of ORF[a]	%ID[b]	Viral group[c]	Gene	Depth[d]
Pepperwood Preserve	*O. sierrensis*	32	1011	34.9	dsRNA	RdRp[e]	123
		32	1012	21.4	dsRNA	PArp[f]	123
		13869	151	27.1	*Rhabdoviridae*	Glycoprotein	6
Bolinas	*Cx. pipiens*	none					
Stinson Beach	*Cx. pipiens*	none					
San Rafael	*Cx. pipiens*	30	2435	26.5	*Bunyaviridae*	RdRp	2985
		2643	1006	24.6	*Narnavirus*	RdRp	3815
		32851	101	40.4	*Rhabdoviridae*	Nucleocapsid	3
Mill Valley	*C. incidens*	84	981	32.8	dsRNA	RdRp	45
		84	975	20.9	dsRNA	PArp	45
		89	917	35.4	dsRNA	RdRp	67
		89	1031	23.9	dsRNA	PArp	67
San Francisco	*Cx. pipiens*	84	2371	26.5	*Bunyaviridae*	RdRp	2503
		1516	1027	24.2	*Narnavirus*	RdRp	6199
		6587	408	33.1	*Rhabdoviridae*	Nucleocapsid	31
San Mateo	*C. incidens*	119	981	33.0	dsRNA	RdRp	140
		119	975	20.9	dsRNA	PArp	140
		1366	695	35.5	dsRNA	RdRp	10
		1366	395	23.9	dsRNA	PArp	10
		20718	275	34.5	dsRNA	RdRp	7
		26686	217	28.5	dsRNA	RdRp	6

[a] *In amino acids.*
[b] *To nearest match within the custom blastp database (see Section "Materials and Methods" and supplementary data at http://dx.doi.org/10.6084/m9.figshare.1247641).*
[c] *As determined by nearest blastp match and phylogenetic reconstruction.*
[d] *As determined using by mapping the quality-checked reads back to the assembled contigs using Bowtie 2. Depth is the average coverage across the entire length of the contig. Coverage maps for each contig are available at http://dx.doi.org/10.6084/m9.figshare.1247641*
[e] *RNA-dependent RNA polymerase (RdRp).*
[f] *Proline-alanine rich protein (PArp).*

identification was verified by sequence identity for cytochrome oxidase I (COI) using a leg from the voucher specimen.

DNA Extraction, Library Preparation, and Sequencing

From the several hundred mosquitoes collected, one representative individual female from each of the seven different collection sites was selected for SMS sequencing (**Table 1**). Prior to processing, whole mosquitoes were washed in 70% ethanol, distilled water, and phosphate-buffered saline (PBS) solution to remove external microbes. While we did not explicitly test the final wash for complete removal of external microbes, a similar wash protocol for *Drosophila* was found to be sufficient for this purpose (Chandler et al., 2011). Washed samples were then individually homogenized in PBS with steel Lysing Matrix I beads (MP Biomedicals).

Samples were prepared for sequencing by first undergoing total RNA isolation, rRNA subtraction, reverse transcription, and random amplification into cDNA libraries. For RNA extraction, we used the MasterPure^{TM} Complete DNA and RNA Purification kit according to the manufacturer's protocols. After DNase treatment, pellets were resuspended in 30 μl of TE buffer and 1 μl of RiboGuard^{TM} was added to each tube. Next, we performed ribosomal subtraction using Ribo-Zero^{TM} Gold (magnetic beads, human/rat/mouse kit). While we did not design our experiment to explicitly test the efficacy of Ribo-Zero^{TM} Gold on the removal of mosquito ribosomal

TABLE 3 | Bacterial sequences associated with northern California mosquitoes.

Location	Species	Contig designation	Contig length	Family	F %[a]	Genus	G %[b]
Pepperwood preserve	*O. sierrensis*	1977	1546	Bacillaceae 1	100	*Bacillus*	100
		18821	401	Corynebacteriaceae	100	*Corynebacterium*	100
		1951	1521	Enterobacteriaceae	100	*Escherichia/Shigella*	100
Bolinas	*Cx. pipiens*	9384	632	Pasteurellaceae	100	*Actinobacillus*	93
		12068	527	Anaplasmataceae[c,d]	78	*Anaplasma*[c,d]	76
		20316	353		46		36
		12181	534		26		17
		13969	492	Bacillaceae 1[d]	100	*Bacillus*[d]	100
		18774	409		98		98
		20021	394	Enterobacteriaceae[d]	100	*Escherichia/Shigella*[d]	100
		16291	447		100		99
Stinson Beach	*Cx. pipiens*	2161	1446	Anaplasmataceae[c]	99	*Anaplasma*[c]	97
		2025	1539	Bacillaceae 1	100	*Bacillus*	100
		1856	1515	Enterobacteriaceae	100	*Escherichia/Shigella*	100
San Rafael	*Cx. pipiens*	4009	1373	Anaplasmataceae[c]	91	*Anaplasma*[c]	90
		4901	333[e]	Bacillaceae 1	100	*Bacillus*	100
		29993	328	Enterobacteriaceae	100	*Escherichia/Shigella*	99
Mill Valley	*C. incidens*	9233	365	Bacillaceae 1	100	*Bacillus*	100
		7981	658	Enterobacteriaceae	100	*Escherichia/Shigella*	100
		7622	713	Moraxellaceae[d]	100	*Moraxella*[d]	100
		8692	521		100		100
		11314	369		100		100
		10818	394	Propionibacteriaceae	100	*Propionibacterium*	100
San Francisco	*Cx. pipiens*	6804	1445	Anaplasmataceae[c]	99	*Anaplasma*[c]	98
		5560	1545	Bacillaceae 1	100	*Bacillus*	100
		16616	579	Enterobacteriaceae	100	*Escherichia/Shigella*	98
		12354	936	Moraxellaceae	100	*Moraxella*	100
San Mateo	*C. incidens*	8563	1498	Bacillaceae 1	100	*Bacillus*	100
		8084	1521	Enterobacteriaceae	100	*Escherichia/Shigella*	100
		24207	781	Pseudomonadaceae	100	*Pseudomonas*	100

[a]*RDP Family level confidence.*
[b]*RDP Genus level confidence.*
[c]*Manually confirmed to be Wolbachia by querying to the NCBI non-redundant database.*
[d]*Multiple contigs from the same sample match to the same bacterial genus.*
[e]*Only the bacterial portion of this chimeric contig was used.*

RNA (rRNA), we note that the final percentage of mosquito 28S and 18S rRNA per library was, on average, 29.5% and 2.5%, respectively (data not shown). This was determined using the default settings of Bowtie 2 (Langmead et al., 2009) to map the quality checked reads (see below) to any contigs with a closest match to insect large subunit (LSU) and small subunit (SSU) rRNA.

Samples then underwent random reverse transcription followed by random amplification. Samples were tagged with a molecular ID tag (MID) and pooled equimolar for a single flow cell of sequencing. Sequencing was done on the Illumina HiSeq 2000 at HudsonAlpha Institute for Biotechnology, using an insert size of 100 bp paired-end reads. No water or blank sample was included as a negative control [although this is recommended for future sequencing runs (Salter et al., 2014)]. Raw, unprocessed sequencing reads are available through the NCBI Short Read Database as part of BioProject PRJNA269777.

Sequence Processing and Taxonomy Assignment

An average of 35 million reads were produced per library (details for each library are available in **Table 1**). Sequences were quality checked and assembled using the A5 pipeline with the metagenome flag (Tritt et al., 2012). A5 combines sequence quality control, adapter trimming, and contig assembly. The SGA software package removes low quality reads and corrects sequencing errors[1] and Tagdust removes sequencing adapter contamination (Lassmann et al., 2009). Cleaned sequences are used to build contigs with the IDBA-UD assembler (Peng et al., 2012).

To identify any viral sequences associated with these mosquitoes, contigs were translated into all six frames and any open reading frames (ORFs) longer than 100 amino acids in length were further examined. These ORFs were queried against a custom blast database using the blastp algorithm. We used

[1]https://github.com/jts/sga/tree/master/src

TABLE 4 | Fungal sequences associated with northern California mosquitoes.

Location	Species	Contig designation	Contig length	Family	F %[a]	Genus	G %[b]
Pepperwood Preserve	*O. sierrensis*	15039	460	Ascomycota incertae sedisa[c]	21	*Capnobotryella*	19
Bolinas	*Cx. pipiens*	12631	512	Trichocomaceae	89	*Chromocleista*	63
Stinson Beach	*Cx. pipiens*	13384	473	Trichocomaceae	87	*Aspergillus*	48
		26216	322	Trichocomaceae	88	*Chromocleista*	35
		23528	346	Trichocomaceae	99	*Eurotium*	94
		20005	332	Exobasidiaceae	100	*Exobasidium*	100
		19659	379	Cystofilobasidiaceae[d]	100	*Guehomyces*[d]	100
		13424	302		100		100
		20700	376	Clavulinaceae[d]	100	*Multiclavula*[d]	100
		21025	372		75		61
		6619	772	Tricholomataceae[d]	86	*Mycenella*[d]	86
		7662	701		96		96
		3091	1173	Coriolaceae	100	*Poria*	100
San Rafael	*Cx. pipiens*	13553	673	Davidiellaceae	100	*Cladosporium* complex	83
Mill Valley	*C. incidens*	10362	337	Erysiphaceae	98	*Arthrocladiella*	72
San Francisco	*Cx. pipiens*	24609	548	Mycosphaerellaceae	100	*Cercosporella*	87
		23638	565	Davidiellaceae	100	*Cladosporium* complex	63
		52139	304	Clavicipitaceae	100	*Claviceps*	100
		12695	931	Malasseziaceae	100	*Malassezia*	100
		50818	310	Saccharomycetaceae	100	*Saccharomyces*	100
San Mateo	*C. incidens*	86913	332	Trichocomaceae	74	*Eupenicillium*	44
		61855	433	Helotiales incertae sedis	86	*Tetracladium*	86

[a]*RDP Family level confidence.*
[b]*RDP Genus level confidence.*
[c]*Manually confirmed to belong to the family Cordycipitaceae by querying to the NCBI non-redundant database.*
[d]*Multiple contigs from the same sample match to the same fungal genus.*

an *e*-value cutoff of 1×10^{-3}, which is approximately equivalent, for our database, to 30% similarity over 100 amino acids and 20% similarity over 500 amino acids. This custom database contains the entire NCBI non-redundant protein database, the PhAnToMe phage protein database (Aziz et al., 2012), the NCBI Viral RefSeq Database, and the *Aedes aegypti*, *Anopheles gambiae*, and *Culex quinquefasciatus* protein databases (Megy et al., 2012). The rationale for combining databases was to ensure that viral and phage proteins were present, while simultaneously reducing the false positive rate by the inclusion of all possible non-viral or non-phage proteins. While none of the mosquitoes examined in this study have their complete genomes sequenced, we note that this would not reduce our ability to detect viral sequences; rather this simply increases the false positive rate. All ORFs with a closest match to "virus" or "phage" were manually queried to the NCBI website to confirm their identity. Confirmed viral sequences are available on NCBI through the GenBank accession numbers KP642114 to KP642128.

To identify any bacterial and fungal sequences in the datasets, the non-translated contigs were queried using the blastn algorithm (Altschul et al., 1990) to the SILVA SSU and LSU Reference Databases (Release 111; Quast et al., 2013). Contigs with a closest match to "Bacteria" and "Fungi" when queried to the SSU and LSU databases, respectively, and were longer than the 300 bp, were then submitted to Ribosomal Database Project's (RDP) Classifier for taxonomic assignment (Wang et al., 2007). Contigs above a 90% genus (SSU) or 70% family (LSU) confidence cutoff were considered reliable hits (**Tables 3** and **4**). All intermediate files, include blast results and fasta files of significant hits are available at http://dx.doi.org/10.6084/m9.figshare.1247641.

To identify the mosquito host taxa, the non-translated contigs were queried to a custom-made database of mosquito COI genes. Matches were manually queried to the NCBI website to confirm their identity. Mosquito COI sequences are available on NCBI through the GenBank accession numbers KP293419 to KP293425.

Viral Phylogenetic Analysis and Coverage Estimation

Phylogenetic analysis was performed by comparing the complete and contiguous sequence of each identified viral ORFs to related taxa. Alignments were performed using MAFFT v7.058 and the E-INS-i algorithm (Katoh and Standley, 2013). This alignment algorithm is suitable for sequences that contain multiple conserved regions embedded in long unalignable regions[2]. Bayesian analysis was performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003). The substitution model was determined by allowing MrBayes to sample across the fixed amino acid rate matrices. For each ORF, two independent chains were run for 1,000,000 generations. The resulting average standard deviation of split frequencies is indicated in the caption

[2]http://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html

to each figure. Tracer v1.5.0 was used to confirm the stationarity of log likelihoods (Rambaut et al., 2013) and the first 25% of the 10,000 total trees were discarded. Results were visualized using FigTree v1.4.0 (Rambaut, 2012; **Figures 2–5**). All trees are available at http://dx.doi.org/10.6084/m9.figshare.1247641. In each case, the alignments contain highly divergent regions without homologs in all taxa, which are encoded with gaps to represent missing data. To confirm these gap sections do not significantly affect reconstruction, we repeated all phylogenetic analyses using only sites that contain data from at least 50% of the taxa. Phylogenetic reconstruction using MrBayes on these shortened alignments found very similar topologies as presented in **Figures 2–5**, and did not change any conclusions (phylogenetic trees resulting from shortened alignments available at http://dx.doi.org/10.6084/m9.figshare.1247641).

Coverage was determined by mapping the quality checked reads to the nucleotide sequences corresponding to the viral ORFs using Bowtie 2 and default settings (Langmead et al., 2009).

Results and Discussion

In this study, we used RNA SMS to characterize the microbial communities associated with three mosquito species in northern California: *Culex pipiens*, *Culiseta incidens*, and *Ochlerotatus sierrensis* (previously *Aedes sierrensis*). *Cx. pipiens* (common house mosquito) is a known vector for *West Nile virus* and *St. Louis encephalitis virus* (Farajollahi et al., 2011), *C. incidens* (cool-weather mosquito) is not a major vector for *West Nile virus* or any other disease, and *O. sierrensis* (western treehole mosquito) is a major vector for dog heartworm (Ledesma and Harrington, 2011). Samples were collected from a variety of habitats including sylvatic/wild, rural, suburban, and urban (*sensu* Thongsripong et al., 2013). All three of these species have limited ranges (flight range <5 miles), making them good markers for the respective habitats in which they are found. Only female mosquitoes were used, as female mosquitoes, being hematophagous, are the primary vectors of most human and animal diseases.

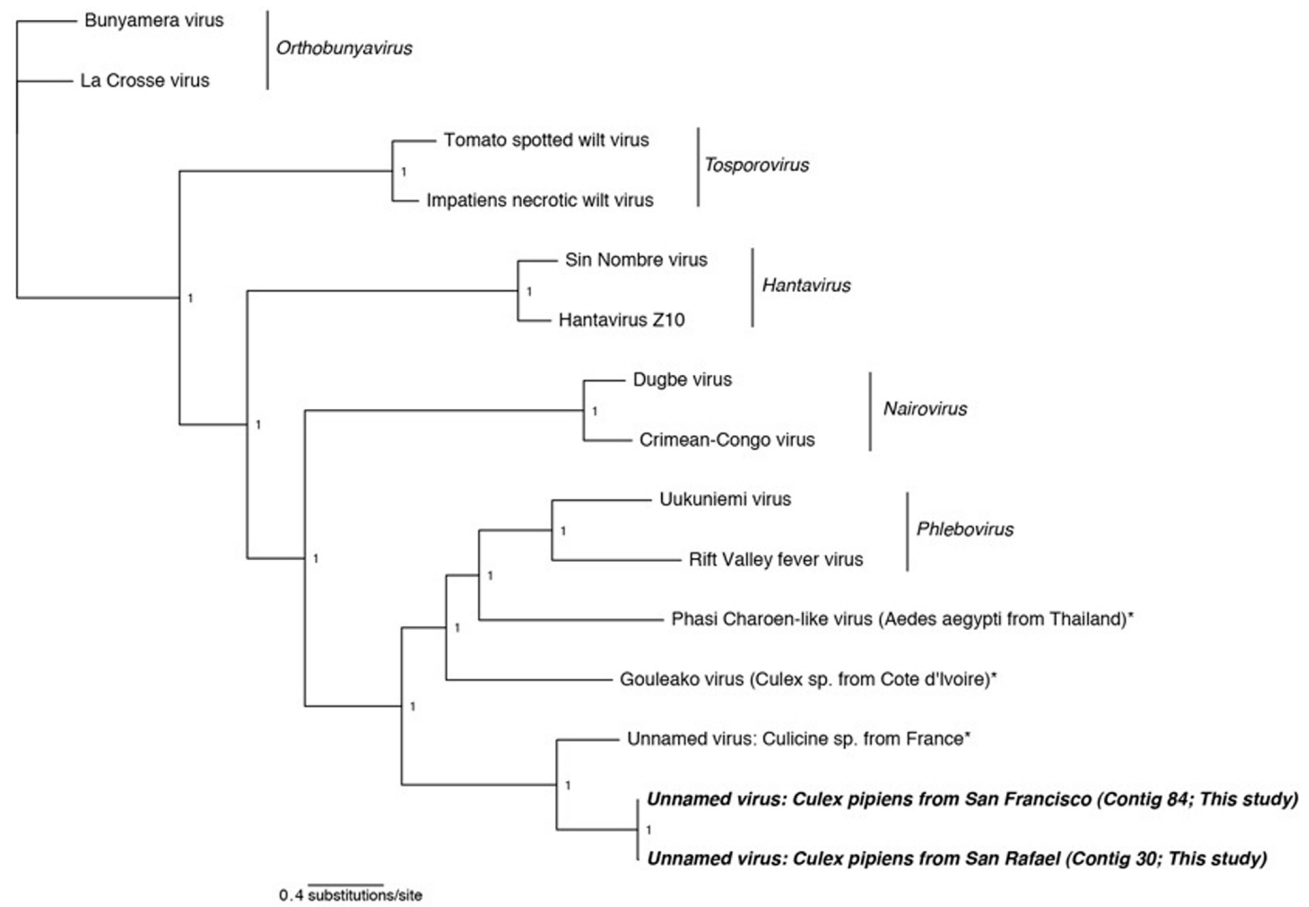


FIGURE 2 | Phylogenetic history of all genera in the *Bunyaviridae* including sequences from this study. This consensus phylogeny was generated in MrBayes (Ronquist and Huelsenbeck, 2003) based on two independent chains run for one million generations to convergence. The average standard deviation of split frequencies was 0.000008. Posterior node probabilities are shown at nodes. Branch lengths are scaled to substitutions/site. Viral sequences uncovered in this study are labeled as such. Closely related mosquito-associated viruses are marked with an asterisk. The accession numbers of all sequences are available at http://dx.doi.org/10.6084/m9.figshare.1247641.

FIGURE 3 | Phylogenetic history of the *Rhabdoviruses* including sequences from this study. This consensus phylogeny was generated in MrBayes (Ronquist and Huelsenbeck, 2003) based on two independent chains run for one million generations to convergence. The average standard deviation of split frequencies was 0.002810. Posterior node probabilities are shown at nodes. Branch lengths are scaled to substitutions/site. Viral sequences uncovered in this study are labeled as such. Closely related mosquito-associated viruses are marked with an asterisk. The accession numbers of all sequences are available at http://dx.doi.org/10.6084/m9.figshare.1247641.

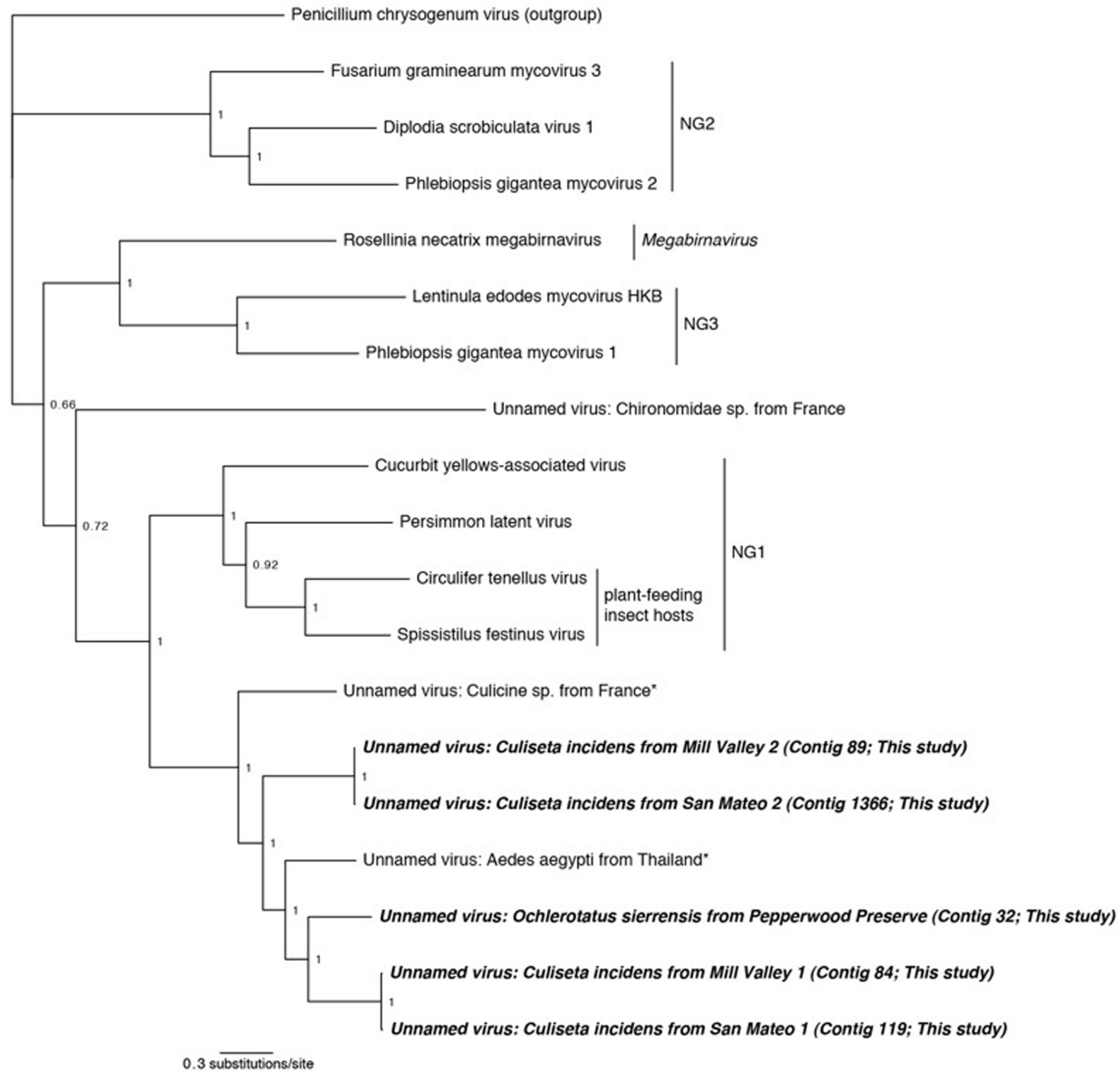


FIGURE 4 | Phylogenetic history of select dsRNA viruses including sequences from this study and Chandler et al. (2014). This consensus phylogeny was generated in MrBayes (Ronquist and Huelsenbeck, 2003) based on two independent chains run for one million generations to convergence. The average standard deviation of split frequencies was 0.000948. Posterior node probabilities are shown at nodes. Branch lengths are scaled to substitutions/site. Viral sequences uncovered in this study are labeled as such. Closely related mosquito-associated viruses are marked with an asterisk. As in Ito et al. (2013), *Penicillium chrysogenum* virus is designated the outgroup. Proposed genera (Ito et al., 2013) are designated NG1, NG2, and NG3. The accession numbers of all sequences are available at http://dx.doi.org/10.6084/m9.figshare.1247641.

As many clinically important mosquito-vectored viruses are RNA based (for example, certain *Alphaviruses*, *Bunyaviridae*, and *Flaviviruses*), we focused on the RNA metagenome of these mosquitoes. Nearly 250 million 100 bp paired-end reads were generated on the Illumina HiSeq platform, roughly evenly distributed among the seven libraries (**Table 1**). After quality control and contig assembly, the identity of the viral sequences associated with these samples was determined by querying the translated contigs against a custom database containing the entire NCBI non-redundant protein database, the PhAnToMe phage protein database (Aziz et al., 2012), the NCBI Viral RefSeq Database, and the *Ae. aegypti*, *An. gambiae*, and *Cx. quinquefasciatus* protein databases (Megy et al., 2012). The identity of the bacterial and fungal sequences associated with these samples was determined by first querying the non-translated contigs to the SILVA SSU and LSU Reference Databases (Quast et al., 2013). Any contigs with a closest match to either bacteria or fungi were then submitted to the RDP Classifier (Wang et al., 2007). Despite using

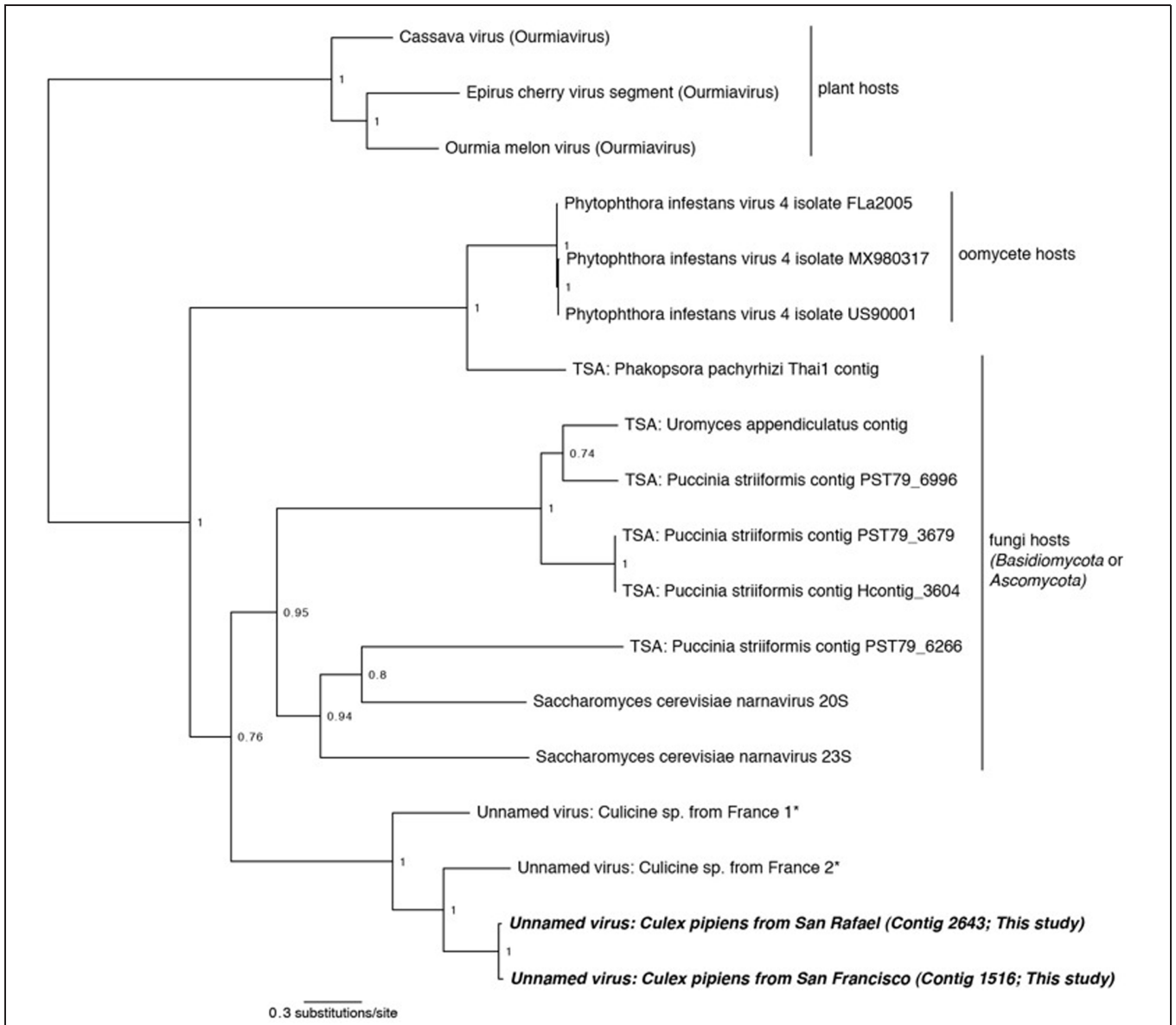


FIGURE 5 | Phylogenetic history of the *Narnaviruses* including sequences from this study. This consensus phylogeny was generated in MrBayes (Ronquist and Huelsenbeck, 2003) based on two independent chains run for one million generations to convergence. *Ourmiaviruses,* which share the RNA-dependent RNA polymerase through a genomic rearrangement, are included as the outgroup. The average standard deviation of split frequencies was 0.004048. Posterior node probabilities are shown at nodes. Branch lengths are scaled to substitutions/site. Viral sequences uncovered in this study are labeled as such. Closely related mosquito-associated viruses are marked with an asterisk. Sequences labeled TSA were uncovered in the NCBI Transcriptome Shotgun Assembly database (Cook et al., 2013). The accession numbers of all sequences are available at http://dx.doi.org/10.6084/m9.figshare.1247641 .

Ribo-Zero^{TM} to deplete vertebrate rRNA, adequate bacterial and fungal RNA remained for microbial classification.

Viral Communities

Consistent with previous studies (Cook et al., 2013; Chandler et al., 2014; Coffey et al., 2014), viral sequences from the lineages *Bunyaviridae*, *Rhabdoviridae*, and *Narnavirus* were uncovered in these mosquitoes (**Table 2**). In addition, we also detected sequences related to an unassigned double-stranded RNA virus (dsRNA) which has been found to be associated with insects (Spear et al., 2010; Cook et al., 2013). No single viral taxon was found in all samples; however, members of the dsRNA group and the *Rhabdoviridae* were detected in three samples. Interestingly, the San Rafael *Cx. pipiens* sample and the San Francisco *Cx. pipiens* sample each contained representatives of the same three viral taxa – *Bunyaviridae*, *Rhabdoviridae*, and *Narnavirus*. Using the methods employed here, viral sequences were not detected in the Stinson Beach or Bolinas *Cx. pipiens* samples. If viruses are present in these samples, they fall below our detection limit (see Section "Materials and Methods").

Bunyaviridae

In two samples, *Cx. pipiens* from San Francisco and *Cx. pipiens* from San Rafael, we uncovered sequences related to the *Bunyaviridae* L segment. There are five recognized *Bunyaviridae* genera (*Phlebovirus, Hantavirus*, *Nairovirus*, *Orthobunyavirus*, and *Tospovirus*), and all except *Hantavirus* are known to be arthropod-vectored. The *Bunyaviridae* have single-stranded negative-sense genomes consisting of three segments. In the genus *Phlebovirus* (which our sequences are most closely related to, see next paragraph), these segments are designated L (encoding the RNA-dependent RNA polymerase), M (encoding the glycoprotein), and S (encoding the nucleocapsid and an ambisense non-structural gene, the NSs, which is necessary for virulence in vertebrates).

Phylogenetic analysis indicates that the detected *Bunyaviridae* sequences are sister to the *Phlebovirus* genus, which includes the agent of Rift Valley fever, and relatives such as *Gouléako virus* (**Figure 2**). This is in congruence with several other mosquito-associated *Bunyaviridae* that have been found in Thailand (Chandler et al., 2014), France (Cook et al., 2013), and West Africa (Marklewitz et al., 2011). Although the two uncovered sequences are more closely related to each other than to any other taxa, these two sequences differ by approximately 6 amino acids substitutions and a 64 amino acid insertion/deletion at the terminal end (alignments available at http://dx.doi.org/10.6084/m9.figshare.1247641).

Attempts to find the M and S *Bunyaviridae* segments using specific blastp and vFam searches (Skewes-Cox et al., 2014) were unsuccessful (data not shown). While this is surprising given the depth of coverage of the L segment (**Table 2**), we note that a previous metagenomic shotgun sequencing study which found a *Bunyaviridae* L segment likewise did not find either the M or the S segments (Cook et al., 2013). The inability to recover these segments may reflect their greater degree of divergence than the L segment, or their reduced representation in the libraries. Targeted amplification of the missing segments by PCR could compensate for the latter, but may have limited success in the case of divergence. Previous work finds that some mosquito viruses lack the NSs gene, which is one of two genes, along with the nucleoprotein, present on the S segment (Marklewitz et al., 2011; Chandler et al., 2014). The absence of the NSs, which is found in all *Phleboviruses* and is necessary for virulence in vertebrates, and the presence of the nucleoprotein, which is required for replication, is suggestive of an arthropod-only lifecycle in these viruses (Marklewitz et al., 2011; Chandler et al., 2014).

The nucleic acid contigs of both *Bunyaviridae* detected here included stop codons on either side of the ORF suggesting that the entire ORF was found. However, the terminal hairpins that are present in all Bunyaviridae were not recovered, despite implementing PRICE (Ruby et al., 2013) to extend the fragments (data not shown). This suggests that the entire L segment was not present in our data.

Rhabdoviridae

In three of our samples, *O. sierrensis* from Pepperwood Preserve, *Cx. pipiens* from San Rafael, and *Cx. pipiens* from San Francisco, we detected sequences related to the family *Rhabdoviridae*. The *Rhabdoviridae* is a diverse family in the order *Mononegavirales* and consists of six established genera (*Vesiculovirus, Lyssavirus, Ephemerovirus, Novirhabdovirus, Cytorhabdovirus,* and *Nucleorhabdovirus*) along with more than 130 unassigned viruses, such as the *Drosophila*-associated *Sigma viruses* (Kuzmin et al., 2009; Longdon et al., 2010). *Rhabdoviridae* are the causative agents of numerous clinically and economically important diseases of humans, livestock, fish, and plants – some of which are arthropod-vectored (Kuzmin et al., 2009). The genome of the *Rhabdoviridae* consists of a single negative-sense RNA segment with, at minimum, five ORFs corresponding to the nucleoprotein, glycoprotein, phosphoprotein, matrix protein, and the RdRp.

Using our methods, we uncovered partial fragments of the *Rhabdoviridae* nucleoprotein (*Cx. pipiens* from San Rafael and *Cx. pipiens* from San Francisco) and the *Rhabdoviridae* glycoprotein (*O. sierrensis* from Pepperwood Preserve). Since the RdRp is the preferred gene for phylogenetic reconstruction (Bourhy et al., 2005; Longdon et al., 2010), we searched for it within these three libraries using specific blastp and vFam searches (Skewes-Cox et al., 2014), but did not recover the RdRp in our datasets. This is unsurprising since the depth of coverage for the three uncovered fragments was very low compared to many of the other uncovered viruses (**Table 2**).

Phylogenetic analysis was performed using MrBayes on concatenated alignments of the nucleoprotein and glycoprotein fragments against reference sequences for these regions (**Figure 3**). The resulting phylogeny successfully recapitulates the relationships between *Rhabdoviridae* lineages ((((*Vesiculovirus, Ephemerovirus), Sigma viruses*), *Lyssavirus), (Cytorhabdovirus, Nucleorhabdovirus*), *Novirhabdovirus*) found by other studies (Longdon et al., 2010; Coffey et al., 2014). Our mosquito-associated *Rhabdoviridae* form a clade with *North Creek virus*, isolated from *Cx. sitiens* in Australia (Coffey et al., 2014) and a unnamed virus isolated from *Cx. tritaeniorhynchus* in Japan (Kuwata et al., 2011). This mosquito-associated clade is sister to the *Vesiculovirus, Ephemerovirus,* and *Sigma virus* clades; basal to these is the mosquito-associated *Moussa virus* (Quan et al., 2010).

Double-Stranded RNA Viruses

In three libraries, we identified sequences that are related to dsRNA viruses previously identified in culicine mosquitoes from France (Cook et al., 2013). These viruses are related to, and have a similar genome structure as, viruses isolated from the plant-feeding insects *Circulifer tenellus* and *Spissistilus festinus* (Spear et al., 2010). The genomes of these viruses consist of a single segment with two ORFs, an RNA-dependent RNA polymerase (RdRp) and a proline–alanine rich protein (PArp) of unknown function. Phylogenetic analysis of the RdRp suggests that the viruses from plant-feeding insects belong to a novel genus of dsRNA viruses (Spear et al., 2010). Phylogenetic analysis using the PArp ORF has not been performed in any previous publication.

We identified sequences related to these dsRNA viruses in three of our samples: *O. sierrensis* from Pepperwood Preserve, *C. incidens* from Mill Valley, and *C. incidens* from San Mateo. Inspired by the discovery of these viral sequences, we searched

a prior dataset from our laboratory [*Ae. aegypti* from Thailand (Chandler et al., 2014)] using the methods described above and identified another dsRNA virus. In several instances, a single contig spanned the entire RdRp and PArp ORFs (contig 32 in the *O. sierrensis* sample from Pepperwood Preserve, contigs 84 and 89 in the *C. incidens* sample from Mill Valley, and contig 119 in the *C. incidens* sample from San Mateo; **Table 2**). In the San Mateo sample, one low coverage contig contains partial sequences of both the RdRp and the PArp (contig 1366) and two additional low coverage contigs (contigs 20718 and 26686) contain fragments of the RdRp that partially overlap with, and are nearly identical to, the contig containing the partial sequences.

Phylogenetic analysis was performed using the complete RdRp ORFs and the partial ORF from the San Mateo sample (contig 1366). Viruses from three proposed novel genera, one *Megabirnavirus*, and *Penicillium chrysogenum* virus were included as comparison taxa (Spear et al., 2010; Ito et al., 2013). Our phylogeny reproduces the overall structure of previous work (Ito et al., 2013) and places all mosquito-associated viruses together (**Figure 4**). Interestingly, both the San Mateo and Mill Valley samples have two distinct dsRNA contigs. This suggests that multiple strains of this dsRNA virus are co-circulating within the northern California *C. incidens* population, a situation similar to that occurring in *Circulifer tenellus* in central California (Spear et al., 2013).

In the dsRNA plant-feeding insect viruses, the RdRp and PArp ORFs are not in frame, but are suspected to be transcribed as a fusion via a -1 ribosomal frameshift just prior to the stop codon of the 5′ (PArp) ORF. The site of this frameshift is predicted to occur at a G_GAA_AAC_stop motif in the virus of French culicine mosquitoes (Cook et al., 2013). In all of our contigs, the PArp and the RdRp were a single base out of frame. In three of our contigs (84 from Mill Valley *C. incidens*, 32 from Pepperwood Preserve *Cx. pipiens*, and 119 from San Mateo *C. incidens*), we identified a G_GAA_AAC_stop motif and in the remaining contigs (89 from Mill Valley *C. incidens* and 1366 from San Mateo *C. incidens*), we identified a G_GGA_AAC_stop motif. This suggests that the two dsRNA ORFs found here are transcribed as a fusion, although enabled by slightly different motifs.

Narnavirus

Two of our samples, *Cx. pipiens* from San Rafael and *Cx. pipiens* from San Francisco, included sequences that are related to the genus *Narnavirus* in the family *Narnaviridae*. *Narnaviruses* have single-stranded positive-sense genomes consisting of a single ORF that encodes an RdRp (King et al., 2012). The type strain *Saccharomyces 20S virus* infects the Ascomycota *Saccharomyces cerevisiae*, while similar sequences have been found associated with Basidiomycota fungi (Cook et al., 2013) and an oomycete plant pathogen (Cai et al., 2013). While the sister genus to *Narnavirus* is the mitochondrial-infecting *Mitovirus*, the RdRp of plant-infecting *Ourmiavirus* is more closely related to *Narnavirus*, as it was acquired via reassortment (Rastgou et al., 2009), and is therefore used as the outgroup in **Figure 5**.

Phylogenetic analysis finds that the *Cx. pipiens Narnaviruses* are more closely related to the other mosquito-associated *Narnaviruses* (Cook et al., 2013) than to any other taxa (**Figure 5**). Since most of the known *Narnaviruses* are suspected to infect Ascomycete or Basdiocymete fungi [the only exception being the oomycete-infecting *Phytophthora infestans RNA virus* (Cai et al., 2013)], we investigated the presence and identity of potential host fungal sequences in our samples (see Section "Fungal Communities"). We also used the methods described in Section "Sequence Processing and Taxonomy Assignment" to identify fungal sequences in the French mosquito dataset that contained two *Narnavirus* sequences (Cook et al., 2013). We did not find any oomycete SSU or LSU sequences in any of the three datasets (data not shown), thus ruling out that the mosquito-associated *Narnaviruses* are in fact associated with the presence of oomycetes.

All three mosquito samples positive for *Narnaviruses* also included sequences of Ascomycete fungi. Although no single fungal genus was common to all three mosquito samples, several [such as *Cladosporium* and *Capnodiales* (see Section "Fungal Communities")] were shared between the two northern California samples. Considering that the mosquito-associated *Narnavirus* clade is sister to *Narnaviruses* of fungi, and that fungal sequences were also detected in the mosquitoes, the *Narnaviruses* from mosquitoes reported here and by Cook et al. (2013) may in fact be instances of virus-infected fungal infections of mosquitoes. Alternatively, these mosquito-associated *Narnaviruses* may represent direct infections of mosquitoes. However, *Saccharomyces 20S RNA virus* is vertically transmitted from mother to daughter cells and horizontally through mating (King et al., 2012), suggesting that host switching may be biologically improbable, particularly from a fungal host to a mosquito host. Taken together, our approach may have revealed an interesting scenario: since the method employed here simultaneously characterized both viruses and fungi within a given sample, it can identify patterns of multi-level nested host-parasite associations. In this case, we have identified putative fungal infecting viruses and the fungi they may be infecting, all in the same hosts. This illustrates the potential utility of our broad approach to microbiota characterization – one that could reveal microbial interactions that may ultimately lead to the development of infectious disease control measures.

Bacteriophages

No phage sequences were uncovered in this dataset, despite including the PhAnToMe phage database in our blastp search. Given that these mosquitoes were associated with a diversity of bacteria (see Section "Bacterial Communities") and that the microbiota of other animals consists of abundant phage (De Paepe et al., 2014), it is reasonable to presume that phages are indeed associated with these mosquitoes as well. There exist two explanations for our inability to find phage associated with these samples. First, our dataset was RNA based and most phages are DNA based. Indeed, known phages of the two most widespread bacteria associated with these samples (*Bacillus* and *Escherichia/Shigella*; see Section "Bacterial Communities") are double-stranded DNA based. Second, perhaps the phages infecting the bacteria within these mosquitoes are undescribed and therefore would not be represented in the databases.

Bacterial Communities

Contigs that could be confidently identified as bacteria using the RDP Classifier were found in all seven samples (**Table 3**). Since there are known limitations of using rRNA as a measurement of either microbial abundance or activity (Blazewicz et al., 2013), the quality-checked reads were not mapped back to these contigs to determine the number of reads used to build each contig. Therefore the data presented here (and of the fungi discussed below) should only be interpreted as representing the presence of that particular taxon and not that taxon's abundance within the community.

Contigs (those that passed the initial 300 bp cutoff) ranged from 328 to 1546 bases in length, which means that some cover nearly the entire 16S gene and all cover at least one hyper-variable region. A single contig with a bacterial match (*Bacillus* in the Mill Valley *C. incidens* sample) was found to be chimeric (data not shown) and only the bacterial portion was used for classification. We identified the following bacterial genera in our samples: *Actinobacillus, Anaplasma* [*Wolbachia* (see below)], *Bacillus, Corynebacterium, Escherichia/Shigella, Haemophilus, Moraxella, Propionibacterium, Pseudomonas* and *Wandonia*. When a genus was found multiple times in the same sample (for example, *Moraxella* in the Mill Valley *C. incidens* sample), the contigs spanned different regions of the 16S gene (see blastn results at http://dx.doi.org/10.6084/m9.figshare.1247641).

Bacillus and *Escherichia/Shigella* were the most widespread genera, as they were found in all seven samples. Notably, a recent review by Minard et al. (2013) concludes that female mosquitoes are mostly colonized by Gammaproteobacteria (of which *Escherichia/Shigella* is a member) and that males are dominated by Firmicutes, such as *Bacillus*. The results of our study, in which only females were investigated, are therefore in partial agreement with this conclusion.

Anaplasma is a genus in the order Rickettsiales. *Wolbachia* is a notable and widespread arthropod-associated bacteria in this order (Hilgenboecker et al., 2008) and we therefore investigated the *Anaplasma* contigs further. Querying these contigs to the NCBI database finds them to be greater than 98% identical to *Wolbachia pipiens* (data not shown). The two samples containing *Wolbachia* were *Cx. pipiens* from San Rafael and Stinson beach. The third *Cx. pipiens* sample, from Bolinas, also had contigs identified as *Anaplasma*, although their confidence was below the 90% cutoff in RDP (**Table 3**). Querying these contigs to NCBI found they were greater than 99% identical to *Wolbachia pipiens* (data not shown, all contigs are available at http://dx.doi.org/10.6084/m9.figshare.1247641). Inspection of the contigs from *C. incidens* and *O. sierrensis* that were shorter than 300 bp or below the 90% confidence cutoff did not reveal any *Anaplasma* or *Wolbachia*-like sequences.

Wolbachia was first described in *Cx. pipiens* mosquitoes (Hertig and Wolbach, 1924). A survey of California mosquitoes (including *Culex, C. incidens,* and *Ochlerotatus*) for *Wolbachia* infections found that, out of 296 individuals and 14 species, only *Cx. pipiens* mosquitoes were infected (Rasgon and Scott, 2004). How *Wolbachia*, which is vertically transmitted, spreads into new taxa is not well understood, but its absence in certain species could be due to inhibition by other members of the mosquito microbiota, as has been shown in *Anopheles* mosquitoes (Hughes et al., 2014). While this study did not find any consistent microbiota differences between *Wolbachia*-infected and *Wolbachia*-free mosquitoes, more comprehensive studies may find such differences that might help account for natural *Wolbachia* distributions.

Of final note, two of the bacterial genera identified, *Propionibacterium* and *Corynebacterium,* are commonly associated with human skin (Grice and Segre, 2011). There are two potential explanations for their presence in these samples. One is that they were acquired from human hosts during a blood meal. Alternatively, these could be contaminants introduced during collection or processing for sequencing. Future studies will include a negative control to identify this possibility (Salter et al., 2014).

Fungal Communities

Contigs that could be confidently classified as fungi using the RDP Classifier were found in all samples except *O. sierrensis* from Pepperwood Preserve (**Table 4**). Manually querying the putative fungal contigs (identified by the initial SILVA database search) from this sample against the NCBI database confidently identified a fungus that was nonetheless below the RDP confidence cutoff (**Table 4**).

Fungal contigs ranged in length from 302 to 1173 bp. Fungi from 13 families were present in our data (**Table 4**). In contrast to the bacterial data, there is no single fungal genus that is present in all samples and only two fungi were found in multiple samples. Specifically, *Chromocleista* was found in *Cx. pipiens* from both Stinson Beach and Bolinas, and *Cladosporium* was found in *Cx. pipiens* from both San Rafael and San Francisco. All the remaining fungal taxa identified were unique to a single sample. More in-depth sequencing may reveal widespread fungi that exist at numbers within the host too low for our methods to detect.

Some of the fungi identified within these samples may not exist in an intimate relationship with mosquitoes. For example, several taxa within the Basidiomycota have a characteristic multicellular stage, such as the *Mycenella* gilled mushrooms identified in the *Cx. pipiens* sample from Stinson Beach. Most likely, dispersing spores of these fungi contaminated water sources used by adult mosquitoes as breeding sites. Similarly, the fungal taxa *Malassezia*, identified in *Cx. pipiens* from San Francisco, is a common human commensal and opportunistic pathogen (Gupta et al., 2004) that likely contaminated the sample through human contact, as with the bacteria discussed above.

Much previous work has focused on using entomopathogenic fungal taxa for mosquito control (Scholte et al., 2004; de Faria and Wraight, 2007). Interestingly, several of the fungi identified belong to families that include many known entomopathogenic fungi including Clavicipitaceae (found in the *Cx. pipiens* sample from San Francisco) and Cordycipitaceae (found in the *O. sierrensis* sample from Pepperwood Preserve). For future mosquito collections, histologic examination would

be necessary to identify visible signs of disease in positive mosquitoes.

This study represents, to our knowledge, the first to use culture-independent techniques to characterize fungal communities in natural populations of mosquitoes. As with other animal hosts [e.g., humans (Huffnagle and Noverr, 2013) and *Drosophila* (Broderick and Lemaitre, 2012)], fungi associated with mosquitoes are relatively understudied compared to bacteria. One study of laboratory-raised *Ae. aegypti* identified *Saccharomyces, Penicillium,* and *Aspergillus* using methods similar to those employed here (Bishop-Lilly et al., 2010). While the current study did find these, or closely related, taxa, more work is needed to fully determine how the fungal communities of laboratory-raised and wild-caught mosquitoes compare.

Mosquito Identification

Five of the seven mosquitoes were identified morphologically or by PCR amplification and Sanger sequencing of the COI gene prior to SMS. For the two specimens that were not identified beforehand, contigs matching mosquito COI and mosquito LSU and SSU were queried against the NCBI database and indicated that these mosquitoes were *Cx. pipiens* and *O. sierrensis* (data not shown, all contigs available at http://dx.doi.org/10.6084/m9.figshare.1247641).

Conclusion

In this study, we used SMS to identify viral, bacterial, and fungal sequences associated with naturally collected mosquitoes. Notably, in all three cases of mosquito-associated viruses (i.e., the *Bunyaviridae*, *Rhabdoviridae*, and dsRNA viruses) the viruses described here were more closely related to other mosquito-associated viruses than to any other viruses. This is despite the fact that the mosquitoes belonged to a variety of taxa (e.g., *Ae. aegypti, C. incidens, O. sierrensis,* and several species of *Culex*) and were collected from a variety of locations (e.g., northern California, Thailand, France, western Africa, and Australia). Furthermore, it should be noted that two of the viral groups discussed here (*Bunyaviridae* and *Rhabdoviridae*) have been extensively studied due to their importance in human and animal health and are therefore well characterized in terms of host breadth. Taken together, a pattern of family level host-specificity (i.e., to the family Culicidae) is emerging. We look forward to future research to sample the viral taxa more extensively and help answer the question of host-specificity and host-switching in the viral tree of life.

With respect to the bacterial and fungal components of the mosquito microbiota, two opposing patterns emerged: several bacterial taxa were widespread regardless of mosquito species or habitat, whereas fungal communities were distinct between individuals even within host species or habitat. These patterns may be an artifact of insufficient sample size and sequencing depth. Much of our sequencing effort was needlessly expended on host rRNA, as over a quarter of the total data can be attributed to the mosquito ribosomal large subunit or small subunit. However, recent techniques have been developed to selectively remove sample-specific rRNA from environmental (Stewart et al., 2010) and mosquito (Kukutla et al., 2013) samples. The use of these techniques will greatly increase the power to detect bacteria, fungi, and viruses in field-caught mosquitoes. Furthermore, many more samples could be combined into a single experiment to allow for the inclusion of multiple replicates per population, host species, or habitat type. Additionally, the physical environment of the mosquitoes, such as larval breeding sites and attractants used in collection, can be sampled. In this way, the complex interactions between host, environment, and all the microbes inhabiting both, can be determined.

Acknowledgments

We thank the California Academy of Sciences' Student Science Fellows high school enrichment program for support, especially Roberta Brett, Sean Edgerton, and Ryan Hulett. This program was generously supported by the Gordon and Betty Moore Foundation, with additional support provided by Mary Austin and Brewster Kahle. V. Piper Kimball and team members of the Marin/Sonoma Mosquito and Vector Control District executed mosquito collections at Stinson Beach, Bolinas, and San Rafael. Durrell Kapan assisted in other mosquito collections. Panpim Thongsripong, Barbara Arrighi, and Anna Sellas provided technical assistance. We thank Jonathan Eisen, Jack Dumbacher, and Joseph DeRisi for use of their servers.

Supplementary Material

The supplemental data associated with this work is available at: http://dx.doi.org/10.6084/m9.figshare.1247641. This includes alignments, tree files, blast outputs, and other intermediate files.

References

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. doi: 10.1016/S0022-2836(05)80360-2

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143–169.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2006). New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl. Environ. Microbiol.* 72, 5734–5741. doi: 10.1128/AEM.00556-06

Aziz, R. K., Devoid, S., Disz, T., Edwards, R. A., Henry, C. S., Olsen, G. J., et al. (2012). SEED servers: high-performance access to the SEED genomes, annotations, and metabolic models. *PLoS ONE* 7:e48053. doi: 10.1371/journal.pone.0048053

Bishop-Lilly, K. A., Turell, M. J., Willner, K. M., Butani, A., Nolan, N. M. E., Lentz, S. M., et al. (2010). Arbovirus detection in insect vectors by rapid, high-throughput pyrosequencing. *PLoS Negl. Trop. Dis.* 4:e878. doi: 10.1371/journal.pntd.0000878

Blazewicz, S. J., Barnard, R. L., Daly, R. A., and Firestone, M. K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. *ISME J.* 7, 2061–2068. doi: 10.1038/ismej.2013.102

Bohart, R. M., and Washino, R. K. (1978). *Mosquitoes of California*, 3rd Edn. Berkeley, CA: University of California, Division of Agricultural Sciences.

Bourhy, H., Cowley, J. A., Larrous, F., Holmes, E. C., and Walker, P. J. (2005). Phylogenetic relationships among rhabdoviruses inferred using the L polymerase gene. *J. Gen. Virol.* 86, 2849–2858. doi: 10.1099/vir.0.81128-0

Broderick, N. A., and Lemaitre, B. (2012). Gut-associated microbes of *Drosophila melanogaster*. *Gut Microbes* 3, 307–321. doi: 10.4161/gmic.19896

Cai, G., Krychiw, J. F., Myers, K., Fry, W. E., and Hillman, B. I. (2013). A new virus from the plant pathogenic oomycete *Phytophthora infestans* with an 8 kb dsRNA genome: the sixth member of a proposed new virus genus. *Virology* 435, 341–349. doi: 10.1016/j.virol.2012.10.012

Chandler, J. A., Lang, J. M., Bhatnagar, S., Eisen, J. A., and Kopp, A. (2011). Bacterial communities of diverse *Drosophila* species: ecological context of a host-microbe model system. *PLoS Genet.* 7:e1002272. doi: 10.1371/journal.pgen.1002272

Chandler, J. A., Thongsripong, P., Green, A., Kittayapong, P., Wilcox, B. A., Schroth, G. P., et al. (2014). Metagenomic shotgun sequencing of a Bunyavirus in wild-caught *Aedes aegypti* from Thailand informs the evolutionary and genomic history of the *Phleboviruses*. *Virology* 464–465, 312–319. doi: 10.1016/j.virol.2014.06.036

Cirimotich, C. M., Ramirez, J. L., and Dimopoulos, G. (2011). Native microbiota shape insect vector competence for human pathogens. *Cell Host Microbe* 10, 307–310. doi: 10.1016/j.chom.2011.09.006

Coffey, L. L., Page, B. L., Greninger, A. L., Herring, B. L., Russell, R. C., Doggett, S. L., et al. (2014). Enhanced arbovirus surveillance with deep sequencing: identification of novel rhabdoviruses and bunyaviruses in Australian mosquitoes. *Virology* 448, 146–158. doi: 10.1016/j.virol.2013.09.026

Cook, S., Chung, B. Y.-W., Bass, D., Moureau, G., Tang, S., McAlister, E., et al. (2013). Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. *PLoS ONE* 8:e80720. doi: 10.1371/journal.pone.0080720

Coon, K. L., Vogel, K. J., Brown, M. R., and Strand, M. R. (2014). Mosquitoes rely on their gut microbiota for development. *Mol. Ecol.* 23, 2727–2739. doi: 10.1111/mec.12771

de Faria, M. R., and Wraight, S. P. (2007). Mycoinsecticides and mycoacaricides: a comprehensive list with worldwide coverage and international classification of formulation types. *Biol. Control* 43, 237–256. doi: 10.1016/j.biocontrol.2007.08.001

De Paepe, M., Leclerc, M., Tinsley, C. R., and Petit, M.-A. (2014). Bacteriophages: an underestimated role in human and animal health? *Front. Cell. Infect. Microbiol.* 4:39. doi: 10.3389/fcimb.2014.00039

Donachie, S. P., Foster, J. S., and Brown, M. V. (2007). Culture clash: challenging the dogma of microbial diversity. *ISME J.* 1, 97–99. doi: 10.1038/ismej.2007.22

Engel, P., Martinson, V. G., and Moran, N. A. (2012). Functional diversity within the simple gut microbiota of the honey bee. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11002–11007. doi: 10.1073/pnas.1202970109

Farajollahi, A., Fonseca, D. M., Kramer, L. D., and Marm Kilpatrick, A. (2011). "Bird biting" mosquitoes and human disease: a review of the role of *Culex pipiens* complex mosquitoes in epidemiology. *Infect. Genet. Evol.* 11, 1577–1585. doi: 10.1016/j.meegid.2011.08.013

Foster, J. A., Bunge, J., Gilbert, J. A., and Moore, J. H. (2012). Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life. *Brief. Bioinform.* 13, 420–429. doi: 10.1093/bib/bbr080

Fukuda, M. M., Klein, T. A., Kochel, T., Quandelacy, T. M., Smith, B. L., Villinski, J., et al. (2011). Malaria and other vector-borne infection surveillance in the U.S. Department of Defense Armed Forces Health Surveillance Center-Global Emerging Infections Surveillance program: review of 2009 accomplishments. *BMC Public Health* 11:S9. doi: 10.1186/1471-2458-11-S2-S9

Ghyselinck, J., Pfeiffer, S., Heylen, K., Sessitsch, A., and De Vos, P. (2013). The effect of primer choice and short read sequences on the outcome of 16S rRNA gene based diversity studies. *PLoS ONE* 8:e71360. doi: 10.1371/journal.pone.0071360

Glaser, R. L., and Meola, M. A. (2010). The native *Wolbachia* endosymbionts of *Drosophila melanogaster* and *Culex quinquefasciatus* increase host resistance to *West Nile virus* infection. *PLoS ONE* 5:e11977. doi: 10.1371/journal.pone.0011977

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol.* 9, 244–253. doi: 10.1038/nrmicro2537

Gupta, A. K., Batra, R., Bluhm, R., Boekhout, T., and Dawson, T. L. (2004). Skin diseases associated with *Malassezia* species. *J. Am. Acad. Dermatol.* 51, 785–798. doi: 10.1016/j.jaad.2003.12.034

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504. doi: 10.1101/gr.112730.110

Hall-Mendelin, S., Allcock, R., Kresoje, N., van den Hurk, A. F., and Warrilow, D. (2013). Detection of arboviruses and other micro-organisms in experimentally infected mosquitoes using massively parallel sequencing. *PLoS ONE* 8:e58026. doi: 10.1371/journal.pone.0058026

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249. doi: 10.1016/S1074-5521(98)90108-9

Hertig, M., and Wolbach, S. B. (1924). Studies on *Rickettsia*-like micro-organisms in insects. *J. Med. Res.* 44, 329–374.7.

Hilgenboecker, K., Hammerstein, P., Schlattmann, P., Telschow, A., and Werren, J. H. (2008). How many species are infected with Wolbachia?–a statistical analysis of current data. *FEMS Microbiol. Lett.* 281, 215–220. doi: 10.1111/j.1574-6968.2008.01110.x

Huffnagle, G. B., and Noverr, M. C. (2013). The emerging world of the fungal microbiome. *Trends Microbiol.* 21, 334–341. doi: 10.1016/j.tim.2013.04.002

Hughes, G. L., Dodson, B. L., Johnson, R. M., Murdock, C. C., Tsujimoto, H., Suzuki, Y., et al. (2014). Native microbiome impedes vertical transmission of *Wolbachia* in *Anopheles* mosquitoes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12498–12503. doi: 10.1073/pnas.1408888111

Ito, T., Suzaki, K., and Nakano, M. (2013). Genetic characterization of novel putative rhabdovirus and dsRNA virus from Japanese persimmon. *J. Gen. Virol.* 94, 1917–1921. doi: 10.1099/vir.0.054445-0

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. doi: 10.1093/molbev/mst010

Kembel, S. W., Wu, M., Eisen, J. A., and Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* 8:e1002743. doi: 10.1371/journal.pcbi.1002743

King, A. M. Q., Adams, M. J., Lefkowitz, E., and Carstens, E. B. (2012). *Virus Taxonomy*. London: Elsevier.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., et al. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* 41, e1. doi: 10.1093/nar/gks808

Kukutla, P., Steritz, M., and Xu, J. (2013). Depletion of ribosomal RNA for mosquito gut metagenomic RNA-seq. *J. Vis. Exp.* e50093. doi: 10.3791/50093

Kuwata, R., Isawa, H., Hoshino, K., Tsuda, Y., Yanase, T., Sasaki, T., et al. (2011). RNA splicing in a new rhabdovirus from *Culex* mosquitoes. *J. Virol.* 85, 6185–6196. doi: 10.1128/JVI.00040-11

Kuzmin, I. V., Novella, I. S., Dietzgen, R. G., Padhi, A., and Rupprecht, C. E. (2009). The rhabdoviruses: biodiversity, phylogenetics, and evolution. *Infect. Genet. Evol.* 9, 541–553. doi: 10.1016/j.meegid.2009.02.005

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi: 10.1186/gb-2009-10-3-r25

Lassmann, T., Hayashizaki, Y., and Daub, C. O. (2009). TagDust–a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25, 2839–2840. doi: 10.1093/bioinformatics/btp527

Ledesma, N., and Harrington, L. (2011). Mosquito vectors of dog heartworm in the United States: vector status and factors influencing transmission efficiency. *Top. Companion Anim. Med.* 26, 178–185. doi: 10.1053/j.tcam.2011.09.005

Longdon, B., Obbard, D. J., and Jiggins, F. M. (2010). Sigma viruses from three species of *Drosophila* form a major new clade in the rhabdovirus phylogeny. *Proc. Biol. Soc.* 277, 35–44. doi: 10.1098/rspb.2009.1472

Ma, M., Huang, Y., Gong, Z., Zhuang, L., Li, C., Yang, H., et al. (2011). Discovery of DNA viruses in wild-caught mosquitoes using small RNA

high throughput sequencing. *PLoS ONE* 6:e24758. doi: 10.1371/journal.pone.0024758

Marklewitz, M., Handrick, S., Grasse, W., Kurth, A., Lukashev, A., Drosten, C., et al. (2011). Gouléako virus isolated from West African mosquitoes constitutes a proposed novel genus in the family *Bunyaviridae*▽. *J. Virol.* 85, 9227–9234. doi: 10.1128/JVI.00230-11

McFall-Ngai, M., Hadfield, M. G., Bosch, T. C. G., Carey, H. V., Domazet-Lošo, T., Douglas, A. E., et al. (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3229–3236. doi: 10.1073/pnas.1218525110

Megy, K., Emrich, S. J., Lawson, D., Campbell, D., Dialynas, E., Hughes, D. S. T., et al. (2012). VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. *Nucleic Acids Res.* 40, D729–D734. doi: 10.1093/nar/gkr1089

Minard, G., Mavingui, P., and Moro, C. V. (2013). Diversity and function of bacterial microbiota in the mosquito holobiont. *Parasit. Vectors* 6, 146. doi: 10.1186/1756-3305-6-146

Morens, D. M., Folkers, G. K., and Fauci, A. S. (2004). The challenge of emerging and re-emerging infectious diseases. *Nature* 430, 242–249. doi: 10.1038/nature02759

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974. doi: 10.1126/science.1198719

Ng, T. F. F., Willner, D. L., Lim, Y. W., Schmieder, R., Chau, B., Nilsson, C., et al. (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* 6:e20579. doi: 10.1371/journal.pone.0020579

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Quan, P.-L., Junglen, S., Tashmukhamedova, A., Conlan, S., Hutchison, S. K., Kurth, A., et al. (2010). Moussa virus: a new member of the Rhabdoviridae family isolated from *Culex decens* mosquitoes in Côte d'Ivoire. *Virus Res.* 147, 17–24. doi: 10.1016/j.virusres.2009.09.013

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi: 10.1093/nar/gks1219

Rambaut, A. (2012). *FigTree v1.4.0*. Available at: http://tree.bio.ed.ac.uk/software/figtree

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J. (2013). *Tracer v1.5.0*. Available at: http://beast.bio.ed.ac.uk/Tracer

Rasgon, J. L., and Scott, T. W. (2004). An initial survey for *Wolbachia* (Rickettsiales: Rickettsiaceae) infections in selected California mosquitoes (Diptera: Culicidae). *J. Med. Entomol.* 41, 255–257. doi: 10.1603/0022-2585-41.2.255

Rastgou, M., Habibi, M. K., Izadpanah, K., Masenga, V., Milne, R. G., Wolf, Y. I., et al. (2009). Molecular characterization of the plant virus genus Ourmiavirus and evidence of inter-kingdom reassortment of viral genome segments as its possible route of origin. *J. Gen. Virol.* 90, 2525–2535. doi: 10.1099/vir.0.013086-0

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Ruby, J. G., Bellare, P., and DeRisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3(Bethesda)* 3, 865–880. doi: 10.1534/g3.113.005967

Runckel, C., Flenniken, M. L., Engel, J. C., Ruby, J. G., Ganem, D., Andino, R., et al. (2011). Temporal analysis of the honey bee microbiome reveals four novel viruses and seasonal prevalence of known viruses, Nosema, and Crithidia. *PLoS ONE* 6:e20656. doi: 10.1371/journal.pone.0020656

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Schmidt, T. M., DeLong, E. F., and Pace, N. R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* 173, 4371–4378.

Scholte, E.-J., Knols, B. G. J., Samson, R. A., and Takken, W. (2004). Entomopathogenic fungi for mosquito control: a review. *J. Insect Sci.* 4, 19. doi: 10.1093/jis/4.1.19

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., and DeRisi, J. L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* 9:e105067. doi: 10.1371/journal.pone.0105067

Spear, A., Sisterson, M. S., Yokomi, R., and Stenger, D. C. (2010). Plant-feeding insects harbor double-stranded RNA viruses encoding a novel proline-alanine rich protein and a polymerase distantly related to that of fungal viruses. *Virology* 404, 304–311. doi: 10.1016/j.virol.2010.05.015

Spear, A., Yokomi, R., French, R., and Stenger, D. C. (2013). Occurrence, sequence polymorphism and population structure of Circulifer tenellus virus 1 in a field population of the beet leafhopper. *Virus Res.* 176, 307–311. doi: 10.1016/j.virusres.2013.06.017

Stewart, F. J., Ottesen, E. A., and DeLong, E. F. (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *ISME J.* 4, 896–907. doi: 10.1038/ismej.2010.18

Thongsripong, P., Green, A., Kittayapong, P., Kapan, D., Wilcox, B., and Bennett, S. (2013). Mosquito vector diversity across habitats in central Thailand endemic for dengue and other arthropod-borne diseases. *PLoS Negl. Trop. Dis.* 7:e2507. doi: 10.1371/journal.pntd.0002507

Tritt, A., Eisen, J. A., Facciotti, M. T., and Darling, A. E. (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS ONE* 7:e42304. doi: 10.1371/journal.pone.0042304

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74. doi: 10.1126/science.1093857

Walker, T., Johnson, P. H., Moreira, L. A., Iturbe-Ormaetxe, I., Frentiu, F. D., McMeniman, C. J., et al. (2011). The wMel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature* 476, 450–453. doi: 10.1038/nature10355

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Wang, S., Ghosh, A. K., Bongio, N., Stebbings, K. A., Lampe, D. J., and Jacobs-Lorena, M. (2012). Fighting malaria with engineered symbiotic bacteria from vector mosquitoes. *Proc. Natl. Acad. Sci. U.S.A.* 109, 12734–12739. doi: 10.1073/pnas.1204158109

Warnecke, F., Luginbühl, P., Ivanova, N., Ghassemian, M., Richardson, T. H., Stege, J. T., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 450, 560–565. doi: 10.1038/nature06269

Weiss, B., and Aksoy, S. (2011). Microbiome influences on insect host vector competence. *Trends Parasitol.* 27, 514–522. doi: 10.1016/j.pt.2011.05.001

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

frontiers in
MICROBIOLOGY

OPINION ARTICLE
published: 12 January 2015
doi: 10.3389/fmicb.2014.00767

Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles

Marilyn J. Roossinck*

Department of Plant Pathology and Environmental Microbiology, Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA, USA
**Correspondence: mjr25@psu.edu*

Edited by:
Bas E. Dutilh, Radboud University Medical Center, Netherlands

Reviewed by:
Hano Maree, Agricultural Research Council, South Africa
Huiquan Liu, Northwest A&F University, China

Keywords: virus ecology, mycoviruses, dsRNA viruses, persistent plant viruses, mutualistic viruses

INTRODUCTION

Most studies of plant viruses have focused on the acute viruses that cause disease in crop and ornamental plants. These viruses are transmitted horizontally, often by insect vectors, and are occasionally transmitted vertically. Although known for at least four decades, the persistent viruses of plants are very poorly studied. These viruses were previously called "cryptic" because they did not appear to illicit any symptoms in infected plants (Boccardo et al., 1987). Persistent plant viruses are not known to be transmitted horizontally, although phylogenetic evidence suggests some level of transmission (Roossinck, 2010). They are vertically transmitted at nearly 100% levels through both ova and pollen (Valverde and Gutierrez, 2007). They have been identified in metagenomic studies by their similarity to known persistent viruses, and because they lack any movement protein, a feature of all known acute viruses that must move through the plant plasmodesmata to establish a systemic infection. Persistent viruses do not move between plant cells, but rather infect every cell and move by cell division.

Most plant persistent viruses have double-stranded (ds) RNA genomes, and encode only an RNA dependent RNA polymerase (RdRp) and a coat protein. Of the well-characterized persistent plant viruses, those in the *Endornaviridae* are the exception. These viruses have a single-stranded (ss) RNA genome, based on their RdRp, and encode a large polyprotein that does not have any apparent coat protein, but encodes a number of additional domains that appear to be derived from diverse sources (Roossinck et al., 2011). They are usually found as dsRNA replicative intermediates.

Viruses of fungi have very similar lifestyles to plant persistent viruses, and several virus families are shared between plants and fungi. Phylogenetic evidence indicates that virus transmission has occurred within and between the two kingdoms (Roossinck, 2010; Roossinck et al., 2011).

Fungal viruses are even less well-studied than plant viruses, and the diversity of these viruses remains mostly unknown. A majority of known fungal viruses have dsRNA genomes, some have ssRNA genomes, and a few examples of DNA viruses are known (Yu et al., 2010). Recently a negative sense ssRNA virus was characterized from a fungus (Liu et al., 2014). Similar to plant viruses, most fungal viruses have been studied in the context of pathogenic fungi. The discovery of the hypovirulence phenotype of *Cryphonectria hypovirus* 1 that suppresses the disease phenotype of the chestnut blight fungus led to a search for other examples that could be exploited to mitigate the effects of plant pathogenic fungi [reviewed in Dawe and Nuss (2013)].

VIRUS DISCOVERY IN PLANTS AND FUNGI

Deep sequencing is proving to be a useful technique for just about everything these days, and the methods have been applied to metagenomic studies of viruses. Unlike studies of other microbes, viruses cannot be analyzed through the use of any universal conserved sequences or motifs, and a variety of techniques have been employed to enrich for viral nucleic acids before sequence analysis. Studies in aquatic viruses have been reported for a number of years (Angly et al., 2006; Labonté and Suttle, 2013). More recently plant viruses have been studied through metagenomics as well (Roossinck, 2012; Stobbe and Roossinck, 2014). A large variety of studies have been done on many different scales, from individual plants to ecosystems. In some studies a single plant species has been targeted, in others a broader sweep is used. These studies are explored in detail in a review by this author and others to be published elsewhere. Here I will explore the discovery of persistent viruses that are extremely common in plants and fungi, but poorly studied, and discuss the implications of these viruses in the ecology of plants and fungi.

Different methods of detection have yielded different levels of persistent viruses. Use of dsRNA-enriched samples yielded very high levels of persistent viruses in plants (Roossinck, 2012). Using the small RNAs involved in plant immunity (siRNAs) has been less successful at detecting many persistent viruses in plants. While the complete sequence of a known endornavirus was assembled with siRNAs (Sela et al., 2012), no novel endornaviruses have been reported with this method. A few siRNAs have been found for partitiviruses, chrysoviruses, and totiviruses, but with very limited genome coverage

(Kreuze, 2014). It is likely that these viruses are not subjected to silencing; with the exception of endornaviruses, they do not expose their dsRNA to the cell, but rather retain their genomes within the virions and extrude only ssRNA into the cytoplasm (Safari and Roossinck, 2014).

Virus discovery in fungi is very limited. Most analyses have been done on fungi of economic importance such as plant pathogenic fungi. A survey of viruses from endophytic fungi derived from two plant species in a wild plant community indicated that the diversity of viruses in this system was much greater than the diversity of fungi, which was in turn much greater than the diversity of plants (Feldman et al., 2012).

In most cases of fungal virus studies, viruses have been discovered from cultured fungi. This eliminates the majority of fungi, which are not culturable (Blackwell, 2011), but have been discovered from environmental samples through specific gene analysis such as ribosomal RNA-related regions and other genes (Seifert, 2009). Traditionally fungi acquired from nature are "purified" by single spore isolation. These practices result in a gross under-estimate of fungal viruses, as many viruses are lost during culture, especially on solid media (unpublished observation), and single spore isolation is a common strategy to obtain cultures "cured" of their viruses. Even though next generation sequencing methods allow for deeper analysis of environmental samples, finding new viruses in fungi without culture is technically challenging. Although some reports have indicated that fungal viruses can be shed into the media when cultured, there is little evidence of extracellular accumulation of most fungal viruses. The lack of any conserved sequences in viruses, such as house-keeping or barcoding genes found in all other life forms, means that sequence-specific primers cannot be used. For viruses of endophytic fungi that have been the focus of fungal virus research in the author's lab, the minimal amount of fungal tissue in plants makes any analysis nearly impossible without culturing the fungus out of the plant. Hence for now we must settle for this very low estimate of fungal virus diversity.

COMMON THEMES FROM PLANTS AND FUNGI

Plants and fungi share several families of viruses. The *Partitviridae* and the *Endornaviridae* are recognized by the International Committee for the Taxonomy of Viruses (King et al., 2012) as infecting both plants and fungi, but biodiversity surveys of plant viruses have also identified members of the *Totiviridae* and *Chrysoviridae* families that traditionally are considered fungal viruses (Roossinck, 2012), and a chrysovirus was recently characterized from radish (Li et al., 2013). In fact viruses from these and related families make up over half of the viruses identified in wild plants (Roossinck, 2012). In plants these viruses appear to maintain a persistent lifestyle (Roossinck, 2010), remaining associated with their hosts for many generations with nearly 100% vertical transmission. Less is known about the lifestyles of fungal viruses. There are few reports of truly acute viruses in fungi. Recently a DNA virus from *Sclerotinia sclerotiorum* was shown to be infectious as a purified virus particle, although it is not clear if this is a mechanism for transmission in nature (Yu et al., 2013). Unlike the plant persistent viruses, fungal viruses can be transmitted between closely related strains of fungus through anastomosis (Milgroom and Hillman, 2011), and evidence of cross-species transmission is apparent in phylogenetic analyses of *Cryphonectria hypovirus* (Liu et al., 2003) and partitiviruses in the *Heterobasdion* (Vainio et al., 2011).

Persistence and high levels of vertical transmission in parasites are correlated with commensal or mutualistic lifestyles (Villarreal, 2007; Márquez and Roossinck, 2012). In some cases we know that persistent viruses are mutualistic (Nakatsukasa-Akune et al., 2005; Márquez et al., 2007). In many cases we don't know enough about their biology to assess their symbiotic lifestyle, but in plants few have any evidence of negative effects on their hosts. For two persistent viruses in *Heterobasidion* species, virus lifestyle was dependent on the fungal environment (Hyder et al., 2013). A complicating factor in understanding the ecology of persistent viruses is that the well-studied persistent viruses are mainly from crop plants, or from economically important fungi; there is virtually no information about any roles these viruses may play in the natural environment where the virus-host relationships evolved.

CONCLUSIONS

The abundance of persistent viruses in plants and fungi imply functions that may contribute to the biology of the host. Unfortunately we have little ecological data about these viruses, and since they often cause no disease they have not been the subject of intensive study. Data-mining from transcriptomic, genomic and metagenomic studies may allow us to address the true ecological role of these viruses. For example, the partitiviruses have poly-A tails, and may be detectable in transcriptome analyses (Jiang et al., 2013). In some cases persistent virus sequences are found integrated into plant or fungal genomes (Liu et al., 2010; Chiba et al., 2011). Deeper analyses along these lines may provide data on time-lines of persistent virus-host relationships.

ACKNOWLEDGMENTS

The author acknowledges the National Science Foundation grant numbers EF-0627108, EPS-0447262, IOS-0950579, and IOS-1157148 and The United States Department of Agriculture grant number OKLR-2007–01012 for previous and continuing research support, and the Pennsylvania State University.

REFERENCES

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Blackwell, M. (2011). The fungi: 1, 2, 3… 5.1 million species? *Am. J. Bot.* 98, 426–438. doi: 10.3732/ajb.1000298

Boccardo, G., Lisa, V., Luisoni, E., and Milne, R. G. (1987). Cryptic plant viruses. *Adv. Virus Res.* 32, 171–214.

Chiba, S., Kondo, H., Tani, A., Saisho, D., Sakamoto, W., Kanematsu, S., et al. (2011). Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathog.* 7:e1002146. doi: 10.1371/journal.ppat.1002146

Dawe, A. L., and Nuss, D. L. (2013). Hypovirus molecular biology: from Koch's postulates to host self-recognition genes that restrict virus transmission. *Adv. Virus Res.* 86, 110–147. doi: 10.1016/B978-0-12-394315-6.00005-2

Feldman, T. S., Morsy, M. R., and Roossinck, M. J. (2012). Are communities of microbial symbionts more diverse than communitiies of macrobial hosts? *Fungal Biol.* 116, 465–477. doi: 10.1016/j.funbio.2012.01.005

Hyder, R., Pennanen, T., Hamberg, L., Vainio, E. J., Piri, T., and Hantula, J. (2013). Two viruses of Heterobasidion confer beneficial, cryptic, or detrimental effects to their hosts in different situations. *Fungal Ecol.* 6, 387–396. doi: 10.1016/j.funeco.2013.05.005

Jiang, L., Wijeratne, A. J., Wijeratne, S., Fraga, M., Meulia, T., Doohan, D., et al. (2013). Profiling mRNAs of two Cuscuta species reveals possible candidate transcripts shared by parasitic plants. *PLoS ONE* 8:e81389. doi: 10.1371/journal.pone.0081389

King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (eds.). (2012). *Virus Taxonomy Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego, CA: Elsevier Academic Press.

Kreuze, J. (2014). "siRNA deep sequencing and assembly: piecing together viral infections," in *Detection and Diagnostics of Plant Pathogens*, eds M. L. Gullino and P. J. M. Bonants (Dordrecht: Springer), 21–38.

Labonté, J. M., and Suttle, C. A. (2013). Previously unknown and highly divergent viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

Li, L., Liu, J., Xu, A., Want, T., Chen, J., and Zhu, X. (2013). Molecular characterization of a trisegmented chrysovirus isolated from the radish *Raphanus sativus*. *Virus Res.* 176, 169–178. doi: 10.1016/j.virusres.2013.06.004

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., et al. (2010). Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *J. Virol.* 84, 11879–11887. doi: 10.1016/j.virusres.2013.06.004

Liu, L., Xie, J., Cheng, J., Fu, Y., Li, G., Yi, X., et al. (2014). Fungal negative-stranded RNA virus that is related to bornavirus and nyavirus. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12205–12210. doi: 10.1073/pnas.1401786111

Liu, Y.-C., Linder-Basso, D., Hillman, B. I., Kaneso, S., and Milgroom, M. G. (2003). Evidence for interspecies transmission of viruses in natural populations of filamentous fungi in the genus Cryphonectria. *Mol. Ecol.* 12, 1619–1628. doi: 10.1046/j.1365-294X.2003.01847.x

Márquez, L. M., Redman, R. S., Rodriguez, R. J., and Roossinck, M. J. (2007). A virus in a fungus in a plant –three way symbiosis required for thermal tolerance. *Science* 315, 513–515. doi: 10.1126/science.1136237

Márquez, L. M., and Roossinck, M. J. (2012). Do persistent RNA viruses fit the trade-off hypothesis of virulence evolution? *Curr. Opin. Virol.* 2, 556–560. doi: 10.1016/j.coviro.2012.06.010

Milgroom, M. G., and Hillman, B. I. (2011). "The ecology and evolution of fungal viruses," in *Studies in Viral Ecology: Microbial and Botanical Host Systems*, ed C. J. Hurst (Hoboken, NJ: John Wiley & Sons), 217–253. doi: 10.1002/9781118025666.ch9

Nakatsukasa-Akune, M., Yamashita, K., Shimoda, Y., Uchiumi, T., Abe, M., Aoki, T., et al. (2005). Suppression of root nodule formation by artificial expression of the TrEnodDR1 (coat protein of White clover cryptic virus 2) gene in *Lotus japonicus*. *Mol. Plant Microbe Interact.* 18, 1069–1080. doi: 10.1094/MPMI-18-1069

Roossinck, M. J. (2010). Lifestyles of plant viruses. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365, 1899–1905. doi: 10.1098/rstb.2010.0057

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annu. Rev. Genet.* 46, 357–367. doi: 10.1146/annurev-genet-110711-155600

Roossinck, M. J., Sabanadzovic, S., Okada, R., and Valverde, R. A. (2011). The remarkable evoluntionary history of endornaviruses. *J. Gen. Virol.* 92, 2674–2678. doi: 10.1099/vir.0.034702-0

Safari, M., and Roossinck, M. J. (2014). How does the genome structure and lifestyle of a virus affect its population variation? *Curr. Opin. Virol.* 9, 39–44. doi: 10.1016/j.coviro.2014.09.004

Seifert, K. A. (2009). Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour.* 9, 83–89. doi: 10.1111/j.1755-0998.2009.02635.x

Sela, N., Luria, N., and Dombrovsky, A. (2012). Genome assembly of Bell pepper endornavirus from small RNA. *J. Virol.* 86, 7721. doi: 10.1128/JVI.00983-12

Stobbe, A. H., and Roossinck, M. J. (2014). Plant virus metagenomics: what we know and why we need to know more. *Front. Plant Sci.* 5:150. doi: 10.3389/fpls.2014.00150

Vainio, E. J., Hakanpää, J., Dai, Y.-C., Hansen, E., Korhonen, K., and Hantula, J. (2011). Species of Heterobasidion host a diverse pool of partitiviruses with global distribution and interspecies transmission. *Fungal Biol.* 115, 1234–1243. doi: 10.1016/j.funbio.2011.08.008

Valverde, R. A., and Gutierrez, D. L. (2007). Transmission of a dsRNA in bell pepper and evidence that it consists of the genome of an endornavirus. *Virus Genes* 35, 399–403. doi: 10.1007/s11262-007-0092-1

Villarreal, L. P. (2007). Virus-host symbiosis mediated by persistence. *Symbiosis* 44, 1–9.

Yu, X., Li, B., Fu, Y., Jiang, D., Ghabrial, S. A., Peng, Y., et al. (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8387–8392. doi: 10.1073/pnas.0913535107

Yu, X., Li, B., Fu, Y., Xie, J., Cheng, J., Ghabrial, S. A., et al. (2013). Extracellular transmission of a DNA mycovirus and its use as a natural fungicide. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1452–1457. doi: 10.1073/pnas.1213755110

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 December 2014; accepted: 16 December 2014; published online: 12 January 2015.
Citation: Roossinck MJ (2015) Metagenomics of plant and fungal viruses reveals an abundance of persistent lifestyles. Front. Microbiol. ***5****:767. doi: 10.3389/fmicb.2014.00767*
This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

MINI REVIEW
published: 13 March 2015
doi: 10.3389/fmicb.2015.00188

Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics

Charlotte J. Houldcroft[1]* *and Judith Breuer*[1,2]

[1] Infection, Inflammation and Rheumatology, Institute of Child Health, University College London, London, UK, [2] Division of Infection and Immunity, University College London, London, UK

Herpesviruses are ubiquitous double-stranded DNA viruses infecting many animals, with the capacity to cause disease in both immunocompetent and immunocompromised hosts. Different herpesviruses have different cell tropisms, and have been detected in a diverse range of tissues and sample types. Metagenomics—encompassing viromics—analyses the nucleic acid of a tissue or other sample in an unbiased manner, making few or no prior assumptions about which viruses may be present in a sample. This approach has successfully discovered a number of novel herpesviruses. Furthermore, metagenomic analysis can identify herpesviruses with high degrees of sequence divergence from known herpesviruses and does not rely upon culturing large quantities of viral material. Metagenomics has had success in two areas of herpesvirus sequencing: firstly, the discovery of novel exogenous and endogenous herpesviruses in primates, bats and cnidarians; and secondly, in characterizing large areas of the genomes of herpesviruses previously only known from small fragments, revealing unexpected diversity. This review will discuss the successes and challenges of using metagenomics to identify novel herpesviruses, and future directions within the field.

Keywords: metagenomics, virus discovery, next-generation sequencing, herpesviruses, *de novo* assembly

OPEN ACCESS

Edited by:
Richard J. Hall, Institute of Environmental Science and Research, New Zealand

Reviewed by:
Marta Canuti, Memorial University of Newfoundland, Canada
Jenny L. Draper, Ministry for Primary Industries, New Zealand

***Correspondence:**
Charlotte J. Houldcroft, Infection, Inflammation and Rheumatology, Institute of Child Health, University College London, Cruciform Building, 30 Guilford Street, London WC1E 6BT, UK
c.houldcroft@ucl.ac.uk

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *03 December 2014*
Paper pending published: *11 January 2015*
Accepted: *20 February 2015*
Published: *13 March 2015*

Citation:
Houldcroft CJ and Breuer J (2015) Tales from the crypt and coral reef: the successes and challenges of identifying new herpesviruses using metagenomics. Front. Microbiol. 6:188. doi: 10.3389/fmicb.2015.00188

Introduction

Herpesviruses are double-stranded DNA viruses which infect many members of the animal kingdom, including mammals, birds and reptiles (McGeoch and Gatherer, 2005), fish (Lepa and Siwicki, 2012), amphibians (Davison et al., 2006), molluscs (Nicolas et al., 1992; Savin et al., 2010), and cnidarians (Vega Thurber et al., 2008; Grasis et al., 2014). They typically infect a large proportion of their target population, spreading through a variety of horizontal and vertical routes. They can cause disease in many settings, e.g., hemorrhage disease in elephants caused by elephant endothelial herpesviruses (Wilkie et al., 2014); cancers caused by the human gammaherpesviruses Epstein-Barr virus (EBV) and Kaposi's sarcoma-associated herpesvirus (KSHV) (Taylor and Blackbourn, 2011); pulmonary infection in cats caused by feline herpesvirus 1 (Thiry et al., 2009) and terrapene herpesvirus 1-associated pneumonia in Eastern box turtles (Sim et al., 2014); seasonal mass mortality in oysters (Nicolas et al., 1992); and herpes simplex virus (HSV) encephalitis in humans (Whitley, 2006). Disease may be associated with primary infection, reactivation of latent infection, immune

Abbreviations: EEHV, elephant endotheliotropic herpesviruses; LCV, lymphocryptovirus.

suppression, or immune senescence. Previous approaches to herpesvirus discovery have utilized a range of methods, discussed in greater depth elsewhere (Bexfield and Kellam, 2011). Briefly, these include: electron microscopy [EBV and cytomegalovirus (CMV); Ho, 2008], PCR-based representational difference analysis (KSHV; Chang et al., 1994), DNA *in situ* hybridization (chelonid herpesviruses Teifke et al., 2000), immunohistochemistry (alcelaphine herpesviruses Klieforth et al., 2002), and polymerase chain reaction (PCR) and Sanger sequencing (primate rhadinoviruses Lacoste et al., 2001, and lizard herpesviruses Wellehan et al., 2004; Literak et al., 2010). Metagenomics is the newest approach to add to this list: a deep-sequencing-based analysis of the nucleic acid contents of a cellular or acellular sample, unreliant on tissue culture and without reference to prior knowledge of which viruses are present in a sample.

Because herpesviruses are widespread among animals, herpesvirus-like sequences are likely to be present in many metagenomic studies. There are also features of herpesvirus molecular biology which increase the likelihood that deep-sequencing studies which are not explicitly metagenomic in nature will nevertheless detect herpesviruses, which may or may not be disease-associated. As widespread DNA viruses, any animal genome sequencing project that uses DNA from primary tissue or samples (e.g., saliva or blood) is likely to include sequences from herpesviruses present in that organism. Even though virus detection is not a primary target of host genome sequencing studies, viruses present in the sample will form a proportion of the sequence reads, and herpesviruses have been discovered in exactly such data (Aswad and Katzourakis, 2014). In this review article, we summarize some of the success stories in identifying novel herpesviruses through metagenomics, and offer a warning on the topic of virus discovery.

Finding the Needle in a Haystack

Much herpesvirus genome sequencing in the past has relied upon large volumes of DNA, acquired through culturing the virus in permissive cell lines (Baer et al., 1984; Lei et al., 2013; Lin et al., 2013). In uncultured or unenriched samples, herpesvirus DNA is present in much smaller proportions than host DNA (Depledge et al., 2011). Herpesvirus genomes are typically hundreds of kilobases in length, complicating the task of Sanger or ultra-deep sequencing of herpesvirus genomes. The problem of poor virus to host DNA ratios applies equally to herpesvirus discovery. It is perhaps unsurprising that many herpesviruses are identified using consensus or degenerate PCR primer sets, but their genomes remain unsequenced (e.g., Bodewes et al., 2014; Sim et al., 2014). The success of identifying novel herpesviruses or of characterizing their genomes further is influenced by the similarity of the novel herpesvirus to related viruses. Some uncharacterized herpesviruses may map well to related herpesviruses, and are relatively easy to assemble from deep-sequence data, or design consensus primers to amplify. Divergent herpesviruses are more difficult to detect and assemble using traditional methods, mapping poorly to existing reference sequences (Pop, 2009). Highly similar and divergent herpesviruses alike may require *de novo* assembly to generate accurate consensus sequences, and potentially further PCR and sequencing to fill gaps in the assembly or confirm novel sequences (e.g., Babra et al., 2012). Metagenomics has an important role in identifying and characterizing divergent herpesviruses, and in allowing the discovery of herpesviruses in hosts or tissue types they might not be expected to be found in.

Deep-Sequencing of Deltaherpesviruses

Elephant endotheliotropic herpesviruses 1A (EEHV1A) and 1B (EEHV1B), causes of hemorrhage disease in African and Asian elephants, required deep sequencing and *de novo* assembly to characterize their complete genome sequences (Wilkie et al., 2013). EEHVs cannot yet be propagated in tissue culture and must be sequenced from primary tissue—in this case, necropsy tissue from a juvenile Asian elephant. While EEHVs were known of from traditional PCR and Sanger-based methods, the complete genome sequence proved to be divergent from known herpesviruses, with 60 novel herpesvirus genes identified in the EEHV1A genome (Ling et al., 2013). Whole genome sequencing of EEHV5 revealed further genetic diversity between strains, with only 60% similarity between EEHV1A and EEHV5 (Wilkie et al., 2014), and 25% nucleotide divergence between coding regions of EEHV1A and EEHV2 (Richman et al., 2014). The use of deep-sequencing to resolve the genome sequences of EEHVs expands our knowledge of herpesvirus diversity, and may lead to the identification of putative deltaherpesviruses in other species.

Herpesviruses Hiding in Plain Sight

Public sequence repositories containing mapped and unmapped reads from large genome-sequencing projects are fertile territory for identifying novel herpesviruses. Untargeted deep and ultra-deep sequencing of mammalian genomes from primary samples can serve as metagenomic datasets for researchers able to re-analyze the data in ways that may not have been anticipated in the original project. Aswad and Katzourakis (2014) screened 14 whole genomes from nine primate sequences available in GenBank for known herpesvirus protein sequence clusters, hoping to identify novel herpesviruses. Characterizing primate herpesvirus diversity is of interest to researchers for a number of reasons: herpesviruses have both zoonotic (Hummeler et al., 1959) and anthroponotic (Huemer et al., 2002) potential in primates, with high fatality rates for humans and non-human primates (Estep et al., 2010); increasing our knowledge of herpesviruses in other species may provide new model systems for understanding how herpesviruses cause disease in humans (Staheli et al., 2014); and the identification of putative endogenous herpesviral elements within primate genomes serves as a fossil record for the herpesviridae (Aswad and Katzourakis, 2014).

Aswad and Katzourakis successfully identified novel herpesvirus-like sequences, and constructed partial herpesvirus genomes, from the whole genome sequences of two primates not previously known to carry herpesviruses: the aye-aye (*Daubentonia madagascariensis*) and Philippine tarsier (*Tarsius syrichta*). Aswad and Katzourakis analyzed both mapped and unmapped reads from the primate genomes, enabling them to identify both exogenous, and chromosomally integrated,

potentially endogenised, herpesviruses. The tarsier herpesvirus sequences were particularly interesting because they are the first report of a potentially endogenous herpesvirus, related to human herpesvirus 6B which chromosomally integrates in a small proportion of humans (Luppi et al., 1993; Tanaka-Taya et al., 2004). The tarsier herpesvirus sequences they reported were heavily mutated, not replication competent and contiguous with stretches of host DNA, supporting the view that this virus has become fully endogenised in tarsier evolutionary history. They were also able to assemble much larger regions of the bonobo (*Pan paniscus*) herpesvirus *Pan paniscus lymphocryptovirus 1* (PpanLCV1), covering an estimated 45% of its genome (∼78,000 bp), which was only previously known from small fragments (∼3000 bp; Ehlers et al., 2003).

While a successful strategy for discovering novel herpesviruses, this approach differs from that employed by metagenomics studies. The authors acted with a reasonable assumption that herpesviruses might be present in the data they were screening, searching only for sequences with similarity to previously reported herpesvirus-like sequences.

Successful identification of novel herpesviruses in host genome sequence datasets also relies on the genomes being sequenced from primary tissue (blood and liver biopsy, in two cases) rather than cell lines, which do not contain the diversity of viruses found in a primary tissue sample. For example, the bonobo genome was sequenced from a bonobo lymphoblastoid cell line which was immortalized in the laboratory using the human herpesvirus EBV (Prufer et al., 2012); this dataset is therefore much more likely to contain human EBV sequences, present in every cell, than primate gammaherpesviruses. Utilizing uncultured primary samples is at the heart of metagenomics.

Going Batty for Metagenomics

Metagenomics may be seen as the opposite side of the coin to host genome sequencing projects: while host genome sequence assembly discards all sequences which are not on target to the host genome, metagenomic studies focus on non-host sequences (reviewed in Mokili et al., 2012). Eliminating or reducing the amount of host nucleic acid present in sequencing data sets has been addressed in some metagenomics studies of bat viral diversity, with success—three recent studies of chiropteran (bat) metagenomes have identified novel herpesvirus sequences, findings confirmed by a fourth study. Bats represent a fifth of the classified mammalian species on earth, and are thought to be a significant reservoir of emerging infectious diseases, particularly those bats found in urban areas or used by humans as a food source (Luis et al., 2013).

Straw-colored fruit bats (*Eidolon helvum*) are widely distributed across Africa and eaten as bushmeat—and therefore a potential zoonotic reservoir (Baker et al., 2013). Studying the viral diversity of this bat is therefore of significant interest to researchers. The sampling approach taken by Baker et al. (2013) was to pool the nucleic acids taken from three sample types: urine, collected non-invasively, and throat and lung swabs obtained from wild, anaesthetized bats from rural and urban locations in Ghana, with sequences assembled *de novo*. They identified sequences from previously unknown bat herpesviruses, with most reads belonging to beta and gammaherpesviruses. The largest proportion of herpesviral reads were obtained from throat swabs, underlining the importance of choosing appropriate primary samples in herpesvirus discovery. The presence of novel herpesviruses was then confirmed by PCR amplification of fragments from the original samples.

Multiple novel herpesviruses were also discovered in a sample of bats from China, which included a broader sampling strategy than in the Baker et al. (2013) study. Anal and pharyngeal swaps were collected from 11 bat species (Wu et al., 2012). In order to maximize the ratio of viral to host/eukaryotic nucleic acid, samples were passed through a filter to remove bacterial and eukaryotic cells. Sequence-independent nucleic acid amplification was used to increase the available genetic material for sequencing, and deep-sequence reads were aligned to the NCBI nt database, identifying novel bat herpesviruses. Even after physical filtering, only 1.2% of sequencing reads were viral, but among these reads, they were able to identify sequences from two previously unknown betaherpesviruses and two gammaherpesviruses. Wu et al. (2012) then used PCR amplification and genome walking to confirm the presence of novel herpesviruses and obtain longer sequences for phylogenetic analysis from the original sample. Clearly low herpesvirus abundance in the initial sample was a factor in preventing whole-genome assembly from the primary material; but metagenomics was integral to identifying the viruses.

A similar study of seven bat species from eastern North American detected a further novel bat betaherpesvirus (Donaldson et al., 2010), from saliva and fresh feces. Samples were passed through a filter to reduce non-viral nucleic acid carry-over, and pooled by species, sex, and age for sequencing to increase the total pool of nucleic acid available. Sequence-independent sample amplification with random hexamers was used, followed by deep-sequencing and *de novo* assembly.

A metagenomic study of French bats (Dacheux et al., 2014) which examined lung, liver and brain tissue from the carcasses of five species was also able to find evidence of the herpesviruses reported by previous studies, reinforcing the ubiquity of herpesviruses among the chiropterans. It is further evident from these studies that saliva, throat swabs and primary tissue samples are integral to herpesvirus discovery.

An Ocean of Herpesviruses

Metagenomics is changing our ideas of where future herpesviruses may be discovered, providing evidence of herpesvirus-like DNA in cnidarians. A metagenomic study of a salt-water coral virome revealed herpesvirus-like sequences in *Porites compressa*. The researchers performed deep-sequencing of the coral in optimal and "stressed" conditions, including increased temperatures, increased acidity, and nutritional stress. They found a small number of herpesvirus-like reads in the coral harvested under optimal conditions, but a much greater number of herpesvirus-like reads were identified in the stressed coral samples. The researchers suggest this is due to lytic reactivation of herpesviruses within the coral under conditions of stress, reflecting what is known

about herpesvirus reactivation in mammals. They were able to PCR amplify and sequence a gene with moderate identity to the thymidylate synthase gene from Herpesvirus saimiri 2 (Vega Thurber et al., 2008).

The same research group followed this up with a metagenomic analysis of four species of coral (*Acropora*, *Diploria*, *Montastraea*, and *Porites*) to test for associations between herpesvirus-like sequence abundance and disease in cnidarians (Soffer et al., 2014), and used transmission electron microscopy to identify herpesvirus-like particles within the cells of healthy coral. Using *de novo* assembly, they found that herpesvirus-like sequences were more common in healthy coral than diseased coral. This result contradicts the previous study, and highlights the difficulties in making quantitative comparisons between metagenomic datasets, when sequences from novel viruses cannot be normalized to gene or genome length for quantification in the manner of RNA-seq data (Mortazavi et al., 2008). While herpesvirus-like sequences had previously been detected in molluscs (e.g., Davison et al., 2005), metagenomics was instrumental in finding herpesvirus-like sequences in cnidarians. Subsequent metagenomic studies of fresh water cnidarians have also found herpesvirus-like sequences in a number of species of hydra (Grasis et al., 2014).

Novel Virus or Novel Contaminant?

There are warnings attached to highly sensitive deep-sequencing and metagenomic discovery of any virus: principally, the risk of contamination. Deep-sequencing and *de novo* assembly of viral nucleic acids provide unprecedented opportunities to identify new viruses, some of which may be pathogenic, but these technologies may also detect contaminating nucleic acids that can be introduced to the sample extraction and sequencing pipeline at many different points.

There have been a number of high profile reports over the last 5 years of "novel" viruses (Xu et al., 2013), which have later been identified as nucleic acid contaminants from commercial sequencing products (e.g., Naccache et al., 2013, 2014; Rosseel et al., 2014). The origins of these nucleic acids are as diverse as mice (Erlwein et al., 2011) and algae (Naccache et al., 2013, 2014). Viral contamination of cell lines (Hue et al., 2010), especially those which have been xenografted during their life in tissue culture (Griffiths et al., 1997, 2002; Paprotka et al., 2011), has also proven to be a significant problem. The story of these "rumor viruses" (Weiss, 2010) must serve as a constant reminder to the virus discovery and metagenomics community. Furthermore, the problem is not confined to viruses, encompassing bacterial contamination of nucleic acid extraction spin columns (Salter et al., 2014) and other laboratory reagents. Screening of control samples and potentially of common laboratory reagents used as each point of the sample preparation process may be necessary for researchers to be confident that they have identified a novel herpesvirus.

The need for further biological characterization to clarify the relationship between a novel pathogen and a disease is a well-recognized problem in viruses studied primarily from molecular data (reviewed in Lipkin, 2010). For example, Bodewes et al. (2014) explicitly comment on the problem of establishing etiology in their report of a novel herpesvirus in harbor and gray seals, initially discovered in juvenile seals with ulcerative gingivitis but also found to be present in a large proportion of healthy controls.

Methodology Influences Success

Metagenomic and deep-sequence analysis of a wide range of eukaryotic organisms over the last decade has shown that herpesviruses are more diverse and ubiquitous than previously imagined. Successful identification of novel herpesviruses or related viruses will rely on the availability of high-quality DNA and RNA from suitable primary samples.

While samples collected non-invasively, such as feces or urine, are immediately appealing for virus discovery because they are easy to collect, they may not be the ideal source material for identifying novel herpesviruses. Metagenomic analysis of bats found most herpesvirus-related sequence reads in throat swab or saliva samples. The primate sequence data in which novel herpesviruses were identified in public sequence repositories were initially sampled from blood and liver. There is an obvious correlation with the cellular tropism of individual herpesvirus species, their abundance in a particular tissue, and the likelihood of discovering a herpesvirus. Herpesviruses with a tropism for lymphocytes may be most readily detectable in blood samples or lymph node biopsies, for example. The physiological and environmental stresses affecting the host may also be an important variable in novel herpesvirus detection.

Metagenomics is an important tool in characterizing herpesviruses previously only known of from sequencing of short PCR amplicons. The success of metagenomics methods improves as the proportion of viral reads within a sample increases. Enriching for total viral material within a sample can be achieved by physical methods early in the sample preparation process, physically reducing the amount of contaminating non-viral nucleic acid using combinations of centrifugation, filtration and nuclease treatments (reviewed in Hall et al., 2014). Further enrichment is then possible, either sequence-independent enrichment using random hexamers or targeted enrichment (such as Sure Select) of specific viruses. All of these approaches may be needed to assemble whole genomes of herpesviruses from uncultured primary samples (Depledge et al., 2011). Targeted enrichment of herpesviruses relies on whole genome sequences being available for the virus of interest. Deep sequencing data from metagenomic studies can provide the sequences needed to design an initial set of PCR primers, as with phocine herpesvirus 7 (Bodewes et al., 2014), or target enrichment baits so that further sequence information can be collected. Metagenomics has also identified novel viruses from many other families, including coronaviruses (Schurch et al., 2014), papillomaviruses (Canuti et al., 2014), and tornoviruses (Ng et al., 2009), to give just a few examples.

Our closest living relatives, chimpanzees, bonobos and gorillas, are a rich source of human herpesvirus-like agents; whole-genome sequences of these viruses would tell us about our shared evolution with these ancient and successful pathogens. The discovery of two previously unknown primate herpesviruses in public sequence

repositories (Aswad and Katzourakis, 2014) makes clear the possibilities for identifying further novel primate herpesviruses through metagenomics in the near future.

There is also the possibility of discovering further human herpesviruses. Humans are unusual among great apes because they have only one rhadinovirus species (KSHV), not two. The presence of three highly divergent forms of KSHV gene K15, including a form found only in southern Africa, has been suggested by some authors (Hayward and Zong, 2007) to indicate that there may be (or may have been) an unknown second KSHV-like rhadinovirus in humans. Deep sequencing and metagenomic analysis of primary tissue samples has been successful in identifying other primate herpesviruses, and as our knowledge of herpesvirus sequence diversity increases, additional human herpesviruses may be identified.

Metagenomics is a promising technique that provides a further tool to identify novel herpesviruses, reducing our reliance on PCR for herpesvirus discovery while also improving our ability to design PCR primers or baits for sequencing studies. Furthermore, metagenomics is increasing our knowledge of the diversity and evolutionary history of herpesviruses; and emphasizing the importance of studying primary samples rather than cell lines or cultured virus isolates if we wish to discover further novel herpesviruses.

Acknowledgments

CJH was funded by GOSH/UCL BRC. JB receives funding from NIHR (www.nihr.co.uk) and UCLH/UCL BRC (http://www.uclhospitals.brc.nihr.ac.uk/).

References

Aswad, A., and Katzourakis, A. (2014). The first endogenous herpesvirus, identified in the tarsier genome, and novel sequences from primate rhadinoviruses and lymphocryptoviruses. *PLoS Genet.* 10:e1004332. doi: 10.1371/journal.pgen.1004332

Babra, B., Watson, G., Xu, W., Jeffrey, B. M., Xu, J. R., Rockey, D. D., et al. (2012). Analysis of the genome of leporid herpesvirus 4. *Virology* 433, 183–191. doi: 10.1016/j.virol.2012.08.002

Baer, R., Bankier, A. T., Biggin, M. D., Deininger, P. L., Farrell, P. J., Gibson, T. J., et al. (1984). DNA sequence and expression of the B95-8 Epstein-Barr virus genome. *Nature* 310, 207–211. doi: 10.1038/310207a0

Baker, K. S., Leggett, R. M., Bexfield, N. H., Alston, M., Daly, G., Todd, S., et al. (2013). Metagenomic study of the viruses of African straw-coloured fruit bats: detection of a chiropteran poxvirus and isolation of a novel adenovirus. *Virology* 441, 95–106. doi: 10.1016/j.virol.2013.03.014

Bexfield, N., and Kellam, P. (2011). Metagenomics and the molecular identification of novel viruses. *Vet. J.* 190, 191–198. doi: 10.1016/j.tvjl.2010.10.014

Bodewes, R., Sanchez Contreras, G. J., Rubio Garcia, A., Hapsari, R., Van De Bildt, M. W., Kuiken, T., et al. (2014). Identification of DNA sequences that imply a novel gammaherpesvirus in seals. *J. Gen. Virol.* doi: 10.1099/vir.0.000029 [Epub ahead of print].

Canuti, M., Deijs, M., Jazaeri Farsani, S. M., Holwerda, M., Jebbink, M. F., De Vries, M., et al. (2014). Metagenomic analysis of a sample from a patient with respiratory tract infection reveals the presence of a gamma-papillomavirus. *Front. Microbiol.* 5:347. doi: 10.3389/fmicb.2014.00347

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M., et al. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 266, 1865–1869. doi: 10.1126/science.7997879

Dacheux, L., Cervantes-Gonzalez, M., Guigon, G., Thiberge, J. M., Vandenbogaert, M., Maufrais, C., et al. (2014). A preliminary study of viral metagenomics of French bat species in contact with humans: identification of new mammalian viruses. *PLoS ONE* 9:e87194. doi: 10.1371/journal.pone.0087194

Davison, A. J., Cunningham, C., Sauerbier, W., and Mckinnell, R. G. (2006). Genome sequences of two frog herpesviruses. *J. Gen. Virol.* 87, 3509–3514. doi: 10.1099/vir.0.82291-0

Davison, A. J., Trus, B. L., Cheng, N., Steven, A. C., Watson, M. S., Cunningham, C., et al. (2005). A novel class of herpesvirus with bivalve hosts. *J. Gen. Virol.* 86, 41–53. doi: 10.1099/vir.0.80382-0

Depledge, D. P., Palser, A. L., Watson, S. J., Lai, I. Y., Gray, E. R., Grant, P., et al. (2011). Specific capture and whole-genome sequencing of viruses from clinical samples. *PLoS ONE* 6:e27805. doi: 10.1371/journal.pone.0027805

Donaldson, E. F., Haskew, A. N., Gates, J. E., Huynh, J., Moore, C. J., and Frieman, M. B. (2010). Metagenomic analysis of the viromes of three North American bat species: viral diversity among different bat species that share a common habitat. *J. Virol.* 84, 13004–13018. doi: 10.1128/JVI.01255-10

Ehlers, B., Ochs, A., Leendertz, F., Goltz, M., Boesch, C., and Matz-Rensing, K. (2003). Novel simian homologues of Epstein-Barr virus. *J. Virol.* 77, 10695–10699. doi: 10.1128/JVI.77.19.10695-10699.2003

Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S., and Mcclure, M. O. (2011). DNA extraction columns contaminated with murine sequences. *PLoS ONE* 6:e23484. doi: 10.1371/journal.pone.0023484

Estep, R. D., Messaoudi, I., and Wong, S. W. (2010). Simian herpesviruses and their risk to humans. *Vaccine* 28(Suppl. 2), B78–B84. doi: 10.1016/j.vaccine.2009.11.026

Grasis, J. A., Lachnit, T., Anton-Erxleben, F., Lim, Y. W., Schmieder, R., Fraune, S., et al. (2014). Species-specific viromes in the ancestral holobiont Hydra. *PLoS ONE* 9:e109952. doi: 10.1371/journal.pone.0109952

Griffiths, D. J., Venables, P. J., Weiss, R. A., and Boyd, M. T. (1997). A novel exogenous retrovirus sequence identified in humans. *J. Virol.* 71, 2866–2872.

Griffiths, D. J., Voisset, C., Venables, P. J., and Weiss, R. A. (2002). Novel endogenous retrovirus in rabbits previously reported as human retrovirus 5. *J. Virol.* 76, 7094–7102. doi: 10.1128/JVI.76.14.7094-7102.2002

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* 195, 194–204. doi: 10.1016/j.jviromet.2013.08.035

Hayward, G. S., and Zong, J. C. (2007). Modern evolutionary history of the human KSHV genome. *Curr. Top. Microbiol. Immunol.* 312, 1–42. doi: 10.1007/978-3-540-34344-8_1

Ho, M. (2008). The history of cytomegalovirus and its diseases. *Med. Microbiol. Immunol.* 197, 65–73. doi: 10.1007/s00430-007-0066-x

Hue, S., Gray, E. R., Gall, A., Katzourakis, A., Tan, C. P., Houldcroft, C. J., et al. (2010). Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology* 7, 111. doi: 10.1186/1742-4690-7-111

Huemer, H. P., Larcher, C., Czedik-Eysenberg, T., Nowotny, N., and Reifinger, M. (2002). Fatal infection of a pet monkey with Human herpesvirus. *Emerg. Infect. Dis.* 8, 639–642. doi: 10.3201/eid0806.010341

Hummeler, K., Davidson, W. L., Henle, W., Laboccetta, A. C., and Ruch, H. G. (1959). Encephalomyelitis due to infection with *Herpesvirus simiae* (herpes B virus); a report of two fatal, laboratory-acquired cases. *N. Engl. J. Med.* 261, 64–68. doi: 10.1056/NEJM195907092610203

Klieforth, R., Maalouf, G., Stalis, I., Terio, K., Janssen, D., and Schrenzel, M. (2002). Malignant catarrhal fever-like disease in Barbary red deer (*Cervus elaphus barbarus*) naturally infected with a virus resembling alcelaphine herpesvirus 2. *J. Clin. Microbiol.* 40, 3381–3390. doi: 10.1128/JCM.40.9.3381-3390.2002

Lacoste, V., Mauclere, P., Dubreuil, G., Lewis, J., Georges-Courbot, M. C., and Gessain, A. (2001). A novel γ2-herpesvirus of the Rhadinovirus 2 lineage in chimpanzees. *Genome Res.* 11, 1511–1519. doi: 10.1101/gr.158601

Lei, H., Li, T., Hung, G. C., Li, B., Tsai, S., and Lo, S. C. (2013). Identification and characterization of EBV genomes in spontaneously immortalized human peripheral blood B lymphocytes by NGS technology. *BMC Genomics* 14:804. doi: 10.1186/1471-2164-14-804

Lepa, A., and Siwicki, A. K. (2012). Fish herpesvirus diseases: a short review of current knowledge. *Acta Vet.* 81, 383–389. doi: 10.2754/avb201281040383

Lin, Z., Wang, X., Strong, M. J., Concha, M., Baddoo, M., Xu, G., et al. (2013). Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *J. Virol.* 87, 1172–1182. doi: 10.1128/JVI.02517-12

Ling, P. D., Reid, J. G., Qin, X., Muzny, D. M., Gibbs, R., Petrosino, J., et al. (2013). Complete genome sequence of elephant endotheliotropic herpesvirus 1A. *Genome Announc.* 1, e0010613. doi: 10.1128/genomeA.00106-13

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377. doi: 10.1128/MMBR.00007-10

Literak, I., Robesova, B., Majlathova, V., Majlath, I., Kulich, P., Fabian, P., et al. (2010). Herpesvirus-associated papillomatosis in a green lizard. *J. Wildl. Dis.* 46, 257–261. doi: 10.7589/0090-3558-46.1.257

Luis, A. D., Hayman, D. T., O'Shea, T. J., Cryan, P. M., Gilbert, A. T., Pulliam, J. R., et al. (2013). A comparison of bats and rodents as reservoirs of zoonotic viruses: are bats special? *Proc. Biol. Sci.* 280, 20122753. doi: 10.1098/rspb.2012.2753

Luppi, M., Marasca, R., Barozzi, P., Ferrari, S., Ceccherini-Nelli, L., Batoni, G., et al. (1993). Three cases of human herpesvirus-6 latent infection: integration of viral genome in peripheral blood mononuclear cell DNA. *J. Med. Virol.* 40, 44–52. doi: 10.1002/jmv.1890400110

McGeoch, D. J., and Gatherer, D. (2005). Integrating reptilian herpesviruses into the family Herpesviridae. *J. Virol.* 79, 725–731. doi: 10.1128/JVI.79.2.725-731.2005

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Mortazavi, A., Williams, B. A., Mccue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Naccache, S. N., Greninger, A. L., Lee, D., Coffey, L. L., Phan, T., Rein-Weston, A., et al. (2013). The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J. Virol.* 87, 11966–11977. doi: 10.1128/JVI.02323-13

Naccache, S. N., Hackett, J. Jr., Delwart, E. L., and Chiu, C. Y. (2014). Concerns over the origin of NIH-CQV, a novel virus discovered in Chinese patients with seronegative hepatitis. *Proc. Natl. Acad. Sci. U.S.A.* 111, E976. doi: 10.1073/pnas.1317064111

Ng, T. F., Manire, C., Borrowman, K., Langer, T., Ehrhart, L., and Breitbart, M. (2009). Discovery of a novel single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics. *J. Virol.* 83, 2500–2509. doi: 10.1128/JVI.01946-08

Nicolas, J., Comps, M., and Cochennec, N. (1992). Herpes-like virus infecting Pacific-oyster larvae, *Crassostrea gigas*. *Bull. Eur. Assoc. Fish Pathol.* 12, 11–13.

Paprotka, T., Delviks-Frankenberry, K. A., Cingoz, O., Martinez, A., Kung, H. J., Tepper, C. G., et al. (2011). Recombinant origin of the retrovirus XMRV. *Science* 333, 97–101. doi: 10.1126/science.1205292

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* 10, 354–366. doi: 10.1093/bib/bbp026

Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., et al. (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486, 527–531. doi: 10.1038/nature11128

Richman, L. K., Zong, J. C., Latimer, E. M., Lock, J., Fleischer, R. C., Heaggans, S. Y., et al. (2014). Elephant endotheliotropic herpesviruses EEHV1A, EEHV1B, and EEHV2 from cases of hemorrhagic disease are highly diverged from other mammalian herpesviruses and may form a new subfamily. *J. Virol.* 88, 13523–13546. doi: 10.1128/JVI.01673-14

Rosseel, T., Pardon, B., De Clercq, K., Ozhelvaci, O., and Van Borm, S. (2014). False-positive results in metagenomic virus discovery: a strong case for follow-up diagnosis. *Transbound. Emerg. Dis.* 61, 293–299. doi: 10.1111/tbed.12251

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1101/007187

Savin, K. W., Cocks, B. G., Wong, F., Sawbridge, T., Cogan, N., Savage, D., et al. (2010). A neurotropic herpesvirus infecting the gastropod, abalone, shares ancestry with oyster herpesvirus and a herpesvirus associated with the amphioxus genome. *Virol. J.* 7, 308. doi: 10.1186/1743-422X-7-308

Schurch, A. C., Schipper, D., Bijl, M. A., Dau, J., Beckmen, K. B., Schapendonk, C. M., et al. (2014). Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS ONE* 9:e105227. doi: 10.1371/journal.pone.0105227

Sim, R. R., Norton, T. M., Bronson, E., Allender, M. C., Stedman, N., Childress, A. L., et al. (2014). Identification of a novel herpesvirus in captive Eastern box turtles (*Terrapene carolina carolina*). *Vet. Microbiol.* 175, 218–223. doi: 10.1016/j.vetmic.2014.11.029

Soffer, N., Brandt, M. E., Correa, A. M., Smith, T. B., and Thurber, R. V. (2014). Potential role of viruses in white plague coral disease. *ISME J.* 8, 271–283. doi: 10.1038/ismej.2013.137

Staheli, J. P., Dyen, M. R., Lewis, P., and Barcy, S. (2014). Discovery and biological characterization of two novel pig-tailed macaque homologs of HHV-6 and HHV-7. *Virology* 471–473C, 126–140. doi: 10.1016/j.virol.2014.10.008

Tanaka-Taya, K., Sashihara, J., Kurahashi, H., Amo, K., Miyagawa, H., Kondo, K., et al. (2004). Human herpesvirus 6 (HHV-6) is transmitted from parent to child in an integrated form and characterization of cases with chromosomally integrated HHV-6 DNA. *J. Med. Virol.* 73, 465–473. doi: 10.1002/jmv.20113

Taylor, G. S., and Blackbourn, D. J. (2011). Infectious agents in human cancers: lessons in immunity and immunomodulation from gammaherpesviruses EBV and KSHV. *Cancer Lett.* 305, 263–278. doi: 10.1016/j.canlet.2010.08.019

Teifke, J. P., Lohr, C. V., Marschang, R. E., Osterrieder, N., and Posthaus, H. (2000). Detection of chelonid herpesvirus DNA by nonradioactive in situ hybridization in tissues from tortoises suffering from stomatitis-rhinitis complex in Europe and North America. *Vet. Pathol.* 37, 377–385. doi: 10.1354/vp.37-5-377

Thiry, E., Addie, D., Belak, S., Boucraut-Baralon, C., Egberink, H., Frymus, T., et al. (2009). Feline herpesvirus infection. ABCD guidelines on prevention and management. *J. Feline Med. Surg.* 11, 547–555. doi: 10.1016/j.jfms.2009.05.003

Vega Thurber, R. L., Barott, K. L., Hall, D., Liu, H., Rodriguez-Mueller, B., Desnues, C., et al. (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18413–18418. doi: 10.1073/pnas.0808985105

Weiss, R. A. (2010). A cautionary tale of virus and disease. *BMC Biol.* 8:124. doi: 10.1186/1741-7007-8-124

Wellehan, J. F., Nichols, D. K., Li, L. L., and Kapur, V. (2004). Three novel herpesviruses associated with stomatitis in Sudan plated lizards (*Gerrhosaurus major*) and a black-lined plated lizard (*Gerrhosaurus nigrolineatus*). *J. Zoo Wildl. Med.* 35, 50–54. doi: 10.1638/03-011

Whitley, R. J. (2006). Herpes simplex encephalitis: adolescents and adults. *Antiviral Res.* 71, 141–148. doi: 10.1016/j.antiviral.2006.04.002

Wilkie, G. S., Davison, A. J., Kerr, K., Stidworthy, M. F., Redrobe, S., Steinbach, F., et al. (2014). First fatality associated with elephant endotheliotropic herpesvirus 5 in an asian elephant: pathological findings and complete viral genome sequence. *Sci. Rep.* 4, 6299. doi: 10.1038/srep06299

Wilkie, G. S., Davison, A. J., Watson, M., Kerr, K., Sanderson, S., Bouts, T., et al. (2013). Complete genome sequences of elephant endotheliotropic herpesviruses 1A and 1B determined directly from fatal cases. *J. Virol.* 87, 6700–6712. doi: 10.1128/JVI.00655-13

Wu, Z., Ren, X., Yang, L., Hu, Y., Yang, J., He, G., et al. (2012). Virome analysis for identification of novel mammalian viruses in bat species from Chinese provinces. *J. Virol.* 86, 10999–11012. doi: 10.1128/JVI.01394-12

Xu, B., Zhi, N., Hu, G., Wan, Z., Zheng, X., Liu, X., et al. (2013). Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10264–10269. doi: 10.1073/pnas.1303744110

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

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 15 December 2014
doi: 10.3389/fmicb.2014.00703

High temporal and spatial diversity in marine RNA viruses implies that they have an important role in mortality and structuring plankton communities

Julia A. Gustavsen[1], Danielle M. Winget[1†], Xi Tian[2] and Curtis A. Suttle[1,3,4]*

[1] *Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada*
[2] *Bioinformatics Graduate Program, Faculty of Science, University of British Columbia, Vancouver, BC, Canada*
[3] *Departments of Botany, and Microbiology & Immunology, University of British Columbia, Vancouver, BC, Canada*
[4] *Canadian Institute for Advanced Research, Toronto, ON, Canada*

Edited by:
Bas E. Dutilh, Radboud University Medical Center, Netherlands

Reviewed by:
Alexander Culley, Université Laval, Canada
Jennifer R. Brum, University of Arizona, USA

****Correspondence:***
Curtis A. Suttle, Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, 2020 - 2207 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: suttle@science.ubc.ca

†Present address:
Danielle M. Winget, Department of Biology, Seattle Pacific University, Seattle, WA, USA

Viruses in the order *Picornavirales* infect eukaryotes, and are widely distributed in coastal waters. Amplicon deep-sequencing of the RNA dependent RNA polymerase (RdRp) revealed diverse and highly uneven communities of picorna-like viruses in the coastal waters of British Columbia (BC), Canada. Almost 300 000 pyrosequence reads revealed 145 operational taxonomic units (OTUs) based on 95% sequence similarity at the amino-acid level. Each sample had between 24 and 71 OTUs and there was little overlap among samples. Phylogenetic analysis revealed that some clades of OTUs were only found at one site; whereas, other clades included OTUs from all sites. Since most of these OTUs are likely from viruses that infect eukaryotic phytoplankton, and viral isolates infecting phytoplankton are strain-specific; each OTU probably arose from the lysis of a specific phytoplankton taxon. Moreover, the patchiness in OTU distribution, and the high turnover of viruses in the mixed layer, implies continuous infection and lysis by RNA viruses of a diverse array of eukaryotic phytoplankton taxa. Hence, these viruses are likely important elements structuring the phytoplankton community, and play a significant role in nutrient cycling and energy transfer.

Keywords: RNA viruses, phytoplankton mortality, viral ecology, pyrosequencing, viral diversity, *Picornavirales*

INTRODUCTION

Viruses are highly abundant and widespread in the oceans (Bergh et al., 1989; Suttle, 2005). Beyond their impacts on host mortality, viruses are significant mediators of biogeochemical processes, horizontal gene transfer, and host community diversity in the oceans (Fuhrman, 1999; Wilhelm and Suttle, 1999; Suttle, 2005). Marine viruses are important pathogens of phytoplankton (Brussaard, 2004), and have been implicated in the termination of blooms (Nagasaki et al., 1994; Schroeder et al., 2003) and with succession in phytoplankton communities (Mühling et al., 2005). Viruses have been characterized that infect a wide variety of phytoplankton such as haptophytes (Bratbak et al., 1993), prasinophytes (Mayer and Taylor, 1979; Brussaard, 2004; Brussaard et al., 2004; Derelle et al., 2008), chlorophytes (Van Etten et al., 1981), diatoms (Shirai et al., 2008), and dinoflagellates (Tomaru et al., 2004). Viruses infecting eukaryotic phytoplankton generally have very narrow host ranges (Short, 2012).

Viruses infecting marine phytoplankton have genomes comprised of double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), double-stranded RNA (dsRNA), and single-stranded RNA (ssRNA) (as reviewed in Short, 2012). Their genomes and particle sizes range from very large dsDNA viruses in the *Phycodnaviridae* to very small ssDNA and ssRNA viruses belonging to the genus *Bacilladnavirus* and order *Picornavirales*, respectively. The order *Picornavirales* is comprised of positive-sense, ssRNA viruses that infect eukaryotes (Le Gall et al., 2008), including ecologically important marine protists. These viruses are small (25–35 nm), icosahedral, and have a conserved genomic organization that includes a replication area comprised of a type III helicase, a 3C-like proteinase, and a type I RNA dependent RNA polymerase (Sanfaçon et al., 2009). Isolates in the *Picornavirales* that are pathogens of marine protists infect a wide diversity of hosts including the bloom-forming raphidophyte *Heterosigma akashiwo* (Tai et al., 2003) (viral family *Marnaviridae*), the thraustochytrid *Aurantiochytrium* sp. (Takao et al., 2006; Yokoyama and Honda, 2007) (proposed viral genus *Labyrnavirus*) and the cosmopolitan diatoms *Rhizosolenia setigera* (Nagasaki et al., 2004) and *Chaetoceros socialis* (Tomaru et al., 2009) (proposed viral genus *Bacillarnavirus*). Viruses in the *Picornavirales* appear to be common and widely distributed in coastal waters (Culley et al., 2003; Culley and Steward, 2007).

Metagenomic and targeted gene studies are uncovering the diversity of marine RNA viruses. For example, phylogenetic analysis of RNA-dependent RNA polymerase (RdRp) sequences from seawater samples supports a monophyletic marine group within the *Picornavirales* (Culley et al., 2003, 2014; Culley and Steward, 2007; Tomaru et al., 2009) and several divergent clades within

this marine group (Culley et al., 2003, 2014; Culley and Steward, 2007). Additionally, metagenomic analyses reveal that there are numerous sequences from aquatic RNA viruses that cannot be assigned to known taxa (Culley et al., 2006, 2014; Djikeng et al., 2009; Steward et al., 2012). Despite the high diversity of marine RNA viruses (Lang et al., 2009), the spatial and temporal distribution of different phylogenetic groups remains unreported, although there is evidence that the taxonomic structure of marine RNA viral communities is highly uneven. For example, in one sample from a metagenomic study from the coastal waters of British Columbia, 59% of the reads assembled into a single contig, while in a second sample 66% of the reads fell into four contigs, with most falling into two genotypes (Culley et al., 2006). However, with only a few 100 reads in total from the two samples, the coverage of the communities was low. Similarly, RNA viral metagenomic data from a freshwater lake (Djikeng et al., 2009) showed little identical sequence overlap among communities, although there was broad taxonomic similarity over time within a location.

Ecological questions about the distribution of marine viruses over time and space have been examined more extensively in bacteriophages, particularly those infecting cyanobacteria. For example, some data reveal no clear patterns of biogeography in cyanophage isolates locally (Clasen et al., 2013), regionally (Jameson et al., 2011) or more globally (Huang et al., 2010). Other data have shown patterns at a regional scale (Marston et al., 2013), although communities in basins that were connected were most similar and those that were separated by land or current boundaries were the least similar. Other data for marine bacteriophages have shown temporal variability (Chen et al., 2009; Wang et al., 2011; Chow and Fuhrman, 2012; Clasen et al., 2013; Marston et al., 2013). If the dynamics of marine bacteriophages and marine RNA viruses are similar, some RNA viral taxa will persist temporally and spatially, while other taxa will be detected sporadically. To test this hypothesis, we examined two samples, taken 5 months apart at the same location, and three samples taken within hours of each other, but 20 km apart in the same coastal basin.

We used high-throughput 454 pyrosequencing to obtain deep coverage of RdRp amplicon sequences and compare the richness of viruses in the *Picornavirales* among samples from the coastal waters of British Columbia, Canada. The results revealed a phylogenetically diverse and spatially variable community of viruses, suggesting that taxon-specific lytic events are important in shaping the phytoplankton community.

MATERIALS AND METHODS

SAMPLING LOCATIONS

To assess viral communities from different coastal habitats, we collected samples from three sites in the Strait of Georgia (49°14.926N 123°35.682W, 49°17.890N 123°43.650W, and 49°23.890N 123°59.706W), and from Jericho Pier (49°16′36.73N, 123°12′05.41W) in British Columbia, Canada) (**Figure 1**). The Strait of Georgia (SOG) is an estuarine-influenced basin that is on average 22 km across, 222 km long and 150 m deep. The upper 50 m of SOG is where most of the variability in physical and

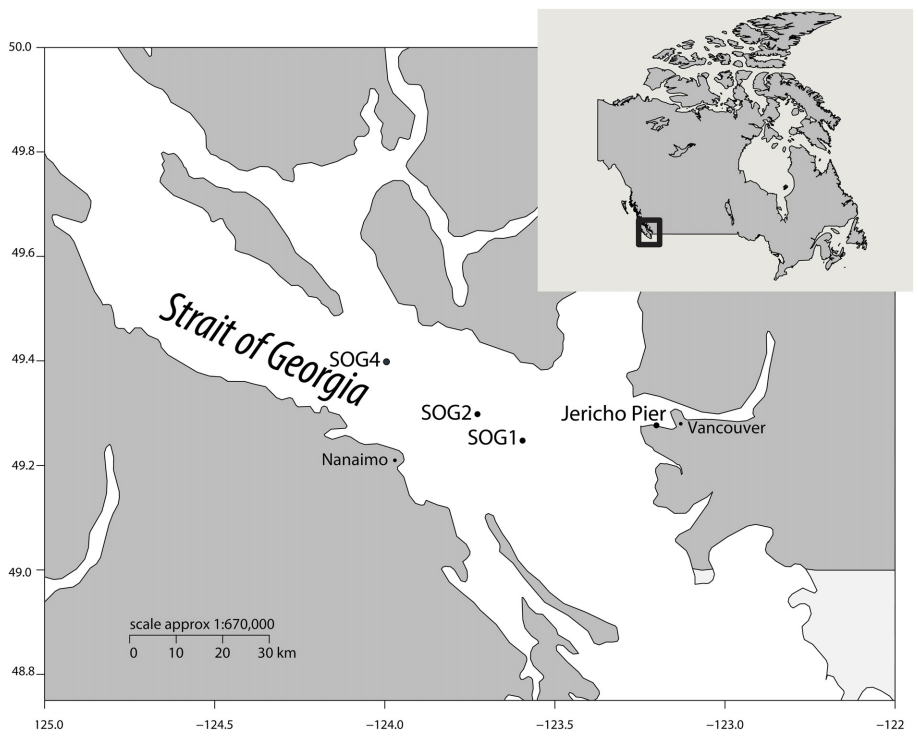


FIGURE 1 | Location of sampling sites. Map showing the location of sampling sites within the Strait of Georgia (SOG) and Jericho Pier, adjacent to Vancouver, British Columbia, Canada. Jericho Pier was sampled in summer (JP-S) and fall (JP-F).

chemical parameters occurs. Jericho Pier (JP) is adjacent to the shoreline, in a well-mixed location with mixed semidiurnal tides.

SAMPLE COLLECTION

On 28 July 2010, a rosette equipped with a Seabird SBE 25 CTD (equipped with Seabird SBE 43 dissolved oxygen sensor, WET Labs WETStar fluorometer, WET Labs C-Star transmissometer, Biospherical Instruments QSP2200PD PAR sensor) and General Oceanics GO-FLO bottles was used to collect 12 L of water from five depths between 2 and 16 m at each of three stations in SOG. The five depths from each station were combined to obtain three integrated water samples. The depths were selected to attempt to encompass the viral diversity at each station, and consisted of a near-surface sample, a sample from the isothermal zone below the mixed layer, and three samples spanning the chlorophyll maximum.

At JP, 60 L of water was pumped from the 1-m depth on 10 July and 12 October 2010. Salinity and temperature at JP were measured using a YSI probe (Yellow Springs, Ohio, USA). For all samples, the water was filtered through 142-mm diameter, 1.2-μm nominal pore-size glass-fiber (GC50 Advantec MFS, Dublin, CA., USA) and 0.22-μm pore-size polyvinyldine (Millipore Bedford, MA, USA) filters. The viral size fraction in the filtrate was concentrated to ∼500 mL (viral concentrate) using tangential flow ultrafiltration using a 30 kDa MW prep-scale Spiral Wound TFF-6 cartridge (Millipore) (Suttle et al., 1991).

Phosphate, silicate and nitrate+nitrite concentrations were determined in duplicate 15-mL seawater samples filtered through 0.45-μm pore-size HA filters (Millipore) and stored at −20°C until air-segmented continuous-flow analysis on a AutoAnalyzer 3 (Bran & Luebbe, Norderstedt Germany). Chlorophyll *a* (Chl *a*) was determined in triplicate by filtering 100 mL of seawater onto 0.45 μm pore-size HA filters (Millipore), and storing the filters in the dark at −20°C until acetone extraction and then analyzed fluorometrically (Parsons et al., 1984). The average and standard error of the replicates was calculated for each sample.

NUCLEIC-ACID EXTRACTION AND PCR

The viral concentrate was filtered twice through 0.22-μm pore-size Durapore PVDF filters (Millipore) in a sterile Sterivex filter unit (Millipore). The filtrate, containing virus-sized particles, was pelleted by ultracentrifugation (Beckman-Coulter, Brea, California, USA) in a SW40 rotor at 108 000 *g* for 5 h at 12°C. The pellet was resuspended overnight in 100 μL of supernatant at 4°C. To digest free DNA, the pellets were incubated with 1U/μl DNAse with a final concentration 5 mM $MgCl_2$ for 3 h at room temperature. Nucleic acids were extracted using a Qiamp Viral Minelute spin kit (Qiagen, Hilden, Germany) according to the manufacturer's directions. To remove DNA, the extracted viral pellets were digested with DNase 1 (amplification grade) (Invitrogen, Carlsbad, California, USA) and the reaction was terminated by adding 2.5 mM EDTA (final concentration) and incubating for 10 min at 65°C. Complementary DNA (cDNA) was generated using Superscript III reverse transcriptase (Invitrogen) with random hexamers (50 ng/μl) as per the manufacturer. PCR was performed using primer set MPL-2 for a targeted set of the marine picornavirus-like RdRp (Culley and Steward, 2007). Each reaction mixture (final volume, 50 μl) consisted of 50 ng of cDNA, 1× (final concentration) PCR buffer (Invitrogen), 2 mM $MgCl_2$, 0.2 mM of each deoxynucleoside triphosphate (Bioline, London, UK), 1 μM of each primer, and 1 U Platinum Taq DNA polymerase. The reaction was run in a PCR Express thermocycler (Hybaid, Ashford, UK) with the following conditions: 94°C for 75 s, followed by 40 cycles of denaturation at 94°C for 45 s, annealing at 43°C for 45 s, and extension at 72°C for 60 s and a final extension step of 9 min at 72°C. PCR products were cleaned using the Minelute PCR purification kit (Qiagen).

LIBRARY PREP AND PYROSEQUENCING

Libraries for each site were prepared for sequencing using NEBNext DNA Library Prep for 454 kit (New England Biolabs, Ipswich, Massachusetts, USA) following the manufacturer's directions, and using Ampure beads (Beckman-Coulter) for size selection and purification using a bead ratio of 0.8:1 beads:library. PCR amplicons were barcoded and sent for 454 Titanium pyrosequencing (Roche, Basel, Switzerland) at the Génome Québec Innovation Centre at the McGill University (Montreal, QC, Canada).

SEQUENCE ANALYSIS

Sequencing reads were quality trimmed using length settings between 100 and 600 bp, with a maximum of 3 primer mismatches for the specific primer, and denoised using the denoiser algorithm in QIIME (version 1.7) with default settings for Titanium data (Caporaso et al., 2010; Reeder and Knight, 2010). Sequences were checked for chimeras using UCHIME (Edgar, 2010) against a nucleotide database of RdRp sequences built using NCBI-BLAST+ by retrieving nucleotide sequences of the RdRp from *Picornavirales* viral isolates (accessed: 8 August 2012), and also using the *denovo* chimera check in UCHIME. The overlap of results from these two methods was defined as chimeric sequences, although none were found in this study. Non-chimeric sequences were queried using BLASTx (Altschul et al., 1990) with an e-value of $1e^{-3}$ against the database of RdRp viral isolates. All sequences with hits were retained and all sequences with no hits were then queried against the non-redundant (nr) Genbank database (Benson et al., 2007) using BLASTx with an e-value of $1e^{-5}$. All the sequences identified as contaminants or as unknown with only 1 read were removed. Remaining sequences were translated to amino acids using FragGeneScan with the 454_10 training option (Rho et al., 2010). Sequences were grouped into operational taxonomic units (OTUs) using UCLUST at a range of similarities from 50 to 100% (Figure S1) using the original seed sequences (centroids) as the output (Edgar, 2010). A similarity of 95% was chosen for this analysis for the following two main reasons: (1) When the NCBI conserved domain alignment for the RdRp region (all *Picornavirales)* was analyzed for percent similarity, the only sequences that displayed greater than 95% similarity in this region were strains of the same virus. (2) The sequences from the control libraries (consisting of 1 clone) clustered into 1 sequence at this percentage (see Supplemental methods). Thus clustering at 95% similarity was a way to use biological and sequence-based information to inform our choice of cut-off to collapse strain level variation and as a conservative

approach to avoid variation that may be present because of the sequencing platform. Control sequences obtained by cloning and Sanger sequencing (see Supplemental Methods) were used to verify the sequence processing methodology. Raw and processed sequence data were deposited in the NCBI BioProject database ID: PRJNA267690.

PHYLOGENETIC ANALYSIS

All OTUs with less than 5 reads were removed and the remaining OTUs were aligned using profile alignment in Muscle (Edgar, 2004) to seed alignments of viral RdRps from the NCBI Conserved Domain Database (Marchler-Bauer et al., 2011). Sequences from other environmental surveys were clustered in the same manner as the reads in this study (using UCLUST at 95%). The clusters are cluster number followed by the Genbank accession numbers contained in that cluster. **0**: 33520549, 33520547, 33520541, 33520533, 33520527, 33520521, 33520519, 33520517, 33520515, 33520513, 33520511, 33520509; **1**: 157280772; **2**: 33520525; **3**: 157280768; **4**: 157280770; **5**: 568801536, 568801534, 568801530, 568801528, 568801510; **6**: 157280786; **7**: 157280774; **8**: 157280780; **9**: 157280788; **10**: 568801494, 568801492, 568801488, 568801482, 568801480, 568801474, 568801470, 568801466, 568801464, 568801462, 568801458; **11**: 157280776; **12**: 157280778; **13**: 568801516; **14**: 568801616, 568801614, 568801606, 568801588, 568801586, 568801584, 568801582, 568801580, 568801574, 568801572, 568801564, 568801550, 568801548, 568801540, 568801532, 568801522, 568801520, 157280784; **15**: 568801508; **16**: 568801542, 568801538, 568801518, 568801486, 568801484, 568801478; **17**: 568801612, 568801610, 568801608, 568801604, 568801602, 568801600, 568801598, 568801596, 568801594, 568801592, 568801590, 568801578, 568801570, 568801568, 568801552, 157280782; **18**: 568801576, 568801566, 568801546, 568801544; **19**: 568801562, 568801556; **20**: 568801526, 568801524, 568801514, 568801512, 568801506, 568801504, 568801502, 568801500, 568801498, 568801496, 568801490, 568801472, 568801460; **21**: 568801476, 568801468; **22**: 568801560, 568801558, 568801554; **23**: 157280744; **24**: 157280758; **25**: 157280748; **26**: 157280746; **27**: 157280742; **28**: 157280766; **29**: 157280756; **30**: 157280762, 157280754, 157280752; **31**: 157280750; **32**: 157280760; **33**: 157280764; **34**: 33520545, 33520543, 33520535, 33520531, 33520529; **35**: 33520539, 33520537, 33520523, 33520507. Alignments were masked using trimAl with the automatic heuristic (Capella-Gutierrez et al., 2009) and edited manually. ProtTest 3.2 was used for amino-acid model selection (Darriba et al., 2011) before building the initial maximum likelihood phylogenetic tree using FastTree (Price et al., 2010). Final trees were done with RA×ML using sequences belonging to viruses in the *Sequiviridae* as the outgroup, and the BLOSUM62 amino-acid model with 100 bootstraps (Stamatakis et al., 2008). The tree was visualized in R (R Core Team, 2014) using the ape package (Paradis, 2012) and edited in Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

STATISTICAL ANALYSIS

Generation of rarefaction curves by random resampling of OTU abundances was performed using the vegan package (Oksanen et al., 2013) in R (R Core Team, 2014). Relative abundances were normalized by randomly resampling 10 000 times using vegan (Oksanen et al., 2013), normalizing to the library with the lowest number of reads and then taking the median. Rank-abundance curves were generated with ggplot2 (Wickham, 2009) using the normalized OTU relative abundances. Scripts used in this project are available as part of QIIME and custom user scripts used to process the data are available here: http://github.com/jooolia/RdRp_454_amplicons_Jericho_and_SOG; doi: 10.5281/zenodo.12509

RESULTS

ENVIRONMENTAL PARAMETERS

The environmental parameters ranged widely among samples (**Figure 2**). Chlorophyll *a* values were lowest at Jericho Pier Fall (JP-F) at 0.16 μg L^{-1} ($\pm$ 0.04) and highest at the Strait of Georgia Station 2 (SOG-2) at 3.2 μg L^{-1} ($\pm$ 0.4). Silicate values for all samples were similar (range of 25.5 to 37.8 μM), except for Jericho Pier Summer (JP-S) when silicate was lower at 6.2 μM ($\pm$ 0.01). Phosphate ranged between 0.90 and 1.3 μM at JP-F, SOG-1, and SOG-4, but was lower at SOG-2 (0.58 μM ($\pm$ 0.03)) and lowest at JP-S (0.06 μM ($\pm$ 0.04)). Nitrate + nitrite values were more variable than the other nutrients and ranged from 1.35 μM ($\pm$ 0.75) at JP-S to 14.6 μM ($\pm$ 0.04) at JP-F. The SOG sites were highly stratified with SOG-4 being the most stratified with a calculated mixed-layer depth of 2 m, while SOG-1 and SOG-2 were similar with a mixed-layer depth of 6 m (**Table 1**).

ANALYSIS OF RdRp SEQUENCES

After quality filtering to remove homopolymers and contaminating reads, 300 180 reads were recovered from the 5 libraries of RdRp amplicons. At all sites the rarefaction curves plateaued indicating that the depth of sampling was adequate to assess the communities (**Figure 3**). From these reads, 265 unique OTUs (at 95% similarity) were identified, including 108 singletons. For further analysis OTUs were excluded that did not contain recognizable RdRp motifs (Koonin, 1991; Le Gall et al., 2008), generally did not align well with other RdRp sequences and those that were not present in any sample after normalization.

Using the above criteria there were 145 OTUs identified in all samples, and between 24 to 71 OTUs per site. The Jericho Pier samples had 116 OTUs of which only 10 (8.6%) were shared between sampling times (**Figure 4**). JP-S had the highest richness with 71 OTUs, of which 59 (83 %) were unique, while JP-F had the second highest richness (49 OTUs), of which 45 (92 %) were unique. The SOG sites together had 64 OTUs, none of which were shared among all sites. SOG-1 and SOG-2 had the lowest number of OTUs (24). SOG-1 had only three OTUs which were unique. However, 21 (33%) were shared between SOG-1 and SOG-4, and 6 (9%) between SOG-2 and SOG-4. The majority of OTUs (75%) from SOG-2 were unique, whereas most OTUs from the other SOG sites were shared with other sites (87% for SOG-1 and 63% for SOG-4).

Rank abundance curves of the viral OTUs showed that at each site most sequences were assigned to only a few OTUs (**Figure 5**).

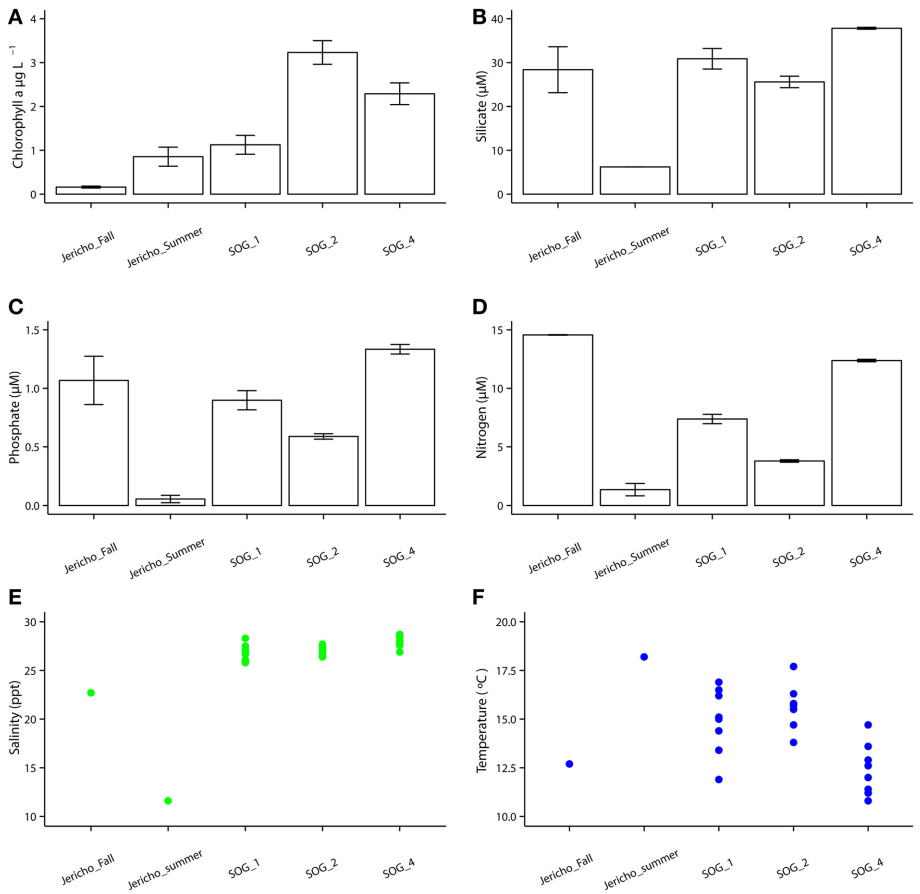


FIGURE 2 | Environmental parameters. (A) Chlorophyll *a* with standard error of the mean from triplicates. **(B)** Silicate with standard error of the mean from duplicates. **(C)** Phosphate with standard error of the mean from duplicates. **(D)** Nitrate+ nitrite with standard error of the mean from duplicates. **(E)** Temperature with each point as one GO-FLO bottle (SOG samples) or from total seawater sample (Jericho samples). **(F)** Salinity with each point representing a seawater sample as one GO-FLO bottle (SOG samples) or from total seawater sample (Jericho samples).

Table 1 | Description of samples and resulting sequencing information.

Location of sampling	Date of sample collection	Latitude Longitude	Reads*	Mixed layer depth (m)
Jericho Pier Summer	10 July 2010	49°16′36.73N, 123°12′05.41W	74,096	–
Jericho Pier Fall	12 October 2010	49°16′36.73N, 123°12′05.41W	84,907	–
SOG 1	28 July 2010	49°14.926N 123°35.682W	55,197	6
SOG 2	28 July 2010	49°17.890N 123°43.650W	12,269	6
SOG 4	28 July 2010	49°23.890N 123°59.706W	73,044	2

**after quality filtering and matching to the RdRp primer set.*

JP-S had the highest richness but the shallowest slope of these curves, demonstrating more evenness in the abundance of OTUs than at the other sites. SOG-4 and JP-S had similar rank abundance curves that were much shallower that those of SOG-1 and SOG-2 (**Figure 5**).

The OTUs that were observed in more than 5 reads were placed in phylogenetic context using a maximum likelihood RA×ML tree (Stamatakis et al., 2008) with sequences from previous RdRp gene surveys and isolated viruses (**Figure 6**). OTUs from this study fell within a well-supported clade that includes all the marine isolates belonging to the *Picornavirales*. Within this group there was a well-supported divide between OTUs grouping in the *Marnaviridae* clade and those grouping with sequences from viruses infecting diatoms and a thraustochytrid. The overall tree topology is not well supported, although there are a number of well-supported clades containing OTUs from this study and other environmental sequences.

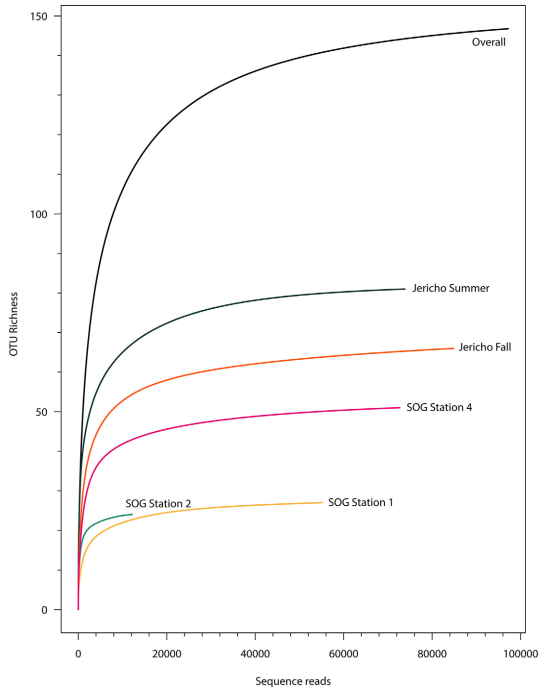


FIGURE 3 | Rarefaction curves. Rarefaction analysis of RdRp amplicons based on Chao1 richness analysis of operational taxonomic units (OTUs) at 95% similarity. Rarefaction curves were resampled using number of reads recovered per library. Rarefaction curves plateau indicating adequate sequencing for these samples.

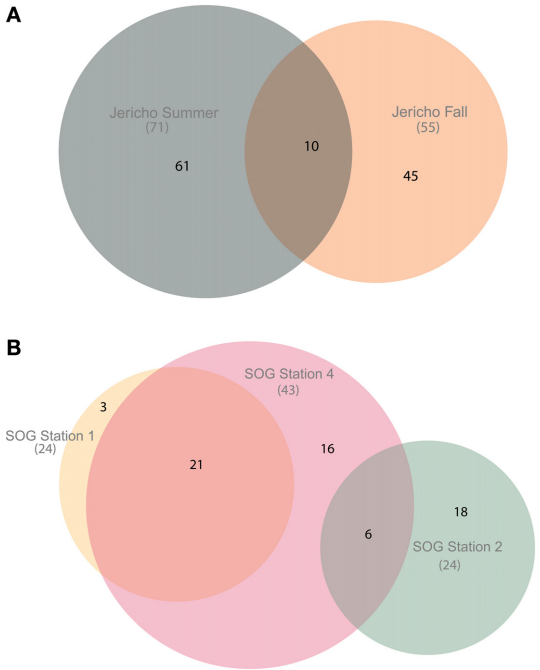


FIGURE 4 | Euler diagrams of normalized RdRp OTUs. (A) Euler diagram of Jericho Pier samples. **(B)** Euler diagram of SOG samples. The OTUs presented were from reads clustered at 95% similarity, comprise only OTUs that could be aligned to the NCBI CDD RdRp alignment, and contained the RdRp motif C. The diagrams were constructed using the venneuler() algorithm (Wilkinson, 2011). The size of the circles is approximately proportional to the number of OTUs recovered per site. The overlap in the diagram describes OTUs that were found at multiple sites and the non-overlapping areas describe OTUs that were unique to that site.

The *Marnaviridae* clade had the greatest number of OTUs (10) associated with it; whereas, very few OTUs (only OTUs 89, 107, 75, and 120) from this study were assigned to clades primarily from Hawaii (Culley and Steward, 2007; Culley et al., 2014). No clade contained OTUs from all sites. The Jericho Pier samples were the most phylogenetically diverse (**Figure 6**, **Table 2**), and contained OTUs (e.g., OTUs 6, 7, 35, 31, 84, 47, 14, 4, 34, 39, 44, 23, 8, 20) that fell into clades that did not contain OTUs from any of the SOG samples. Some clades contained OTUs from both JP-S and JP-F samples; however, many OTUs within the clades were unique to one Jericho Pier sample. Phylogenetic diversity differed among samples, except for OTUs from the two SOG sites with deeper mixed layers, some of which were present in different clades resulting in similar phylogenetic diversity (**Table 2**).

The Strait of Georgia (SOG) sites were sampled within hours of each other, and the water at each site was pooled from multiple depths above, below, and across the chlorophyll maximum. One of the most striking differences among sites was that SOG-1 and SOG -2 had mixed layer depths of 6 m; whereas SOG-4 had a mixed-layer depth of 2 m, and much higher richness and phylogenetic diversity. Sites SOG-1 and SOG-2 had the lowest phylogenetic diversity (**Table 2**). All the OTUs found at SOG-1 (33, 12, 16, 9, 10, 27, 29) were within the *Marnaviridae* clade; similarly, all OTUs (5, 82, 2, 1) from SOG-2 were within one distantly related clade. In both cases OTUs from these clades occurred at SOG-4. SOG-2 did not have the high numbers of HaRNAV-related viruses that were found in all other samples.

DISCUSSION

Pyrosequencing of RdRp gene fragments from coastal samples uncovered much greater genetic diversity than in previous gene surveys (Culley et al., 2003, 2014; Culley and Steward, 2007) and revealed many previously unknown taxonomic groups within the *Picornavirales*. As well, striking differences in the taxonomic richness among samples implies that these viruses infect a wide variety of eukaryotic plankton, but that the mortality imposed on some taxa is highly variable across space and time. Other taxonomic groups within the *Picornavirales* were more widespread, suggesting that infection of some planktonic taxa is more widespread and persistent. These results and their implications are discussed in detail below.

EXPANDING THE KNOWN DIVERSITY OF *PICORNAVIRALES*

The high depth of sequencing and limited diversity in each library (**Figure 3**) gives high confidence that the population structure of RdRp amplicons in each sample has been well characterized (Kemp and Aller, 2004). Although some sequences were closely related to those found in previous studies (Culley et al., 2003; Culley and Steward, 2007) (**Figure 6**), many OTUs formed new clades. Many OTUs were related to *Heterosigma akashiwo* RNA virus (HaRNAV) that infects the toxic bloom-forming raphidophyte *Heterosigma akashiwo* (Tai et al., 2003). HaRNAV is the

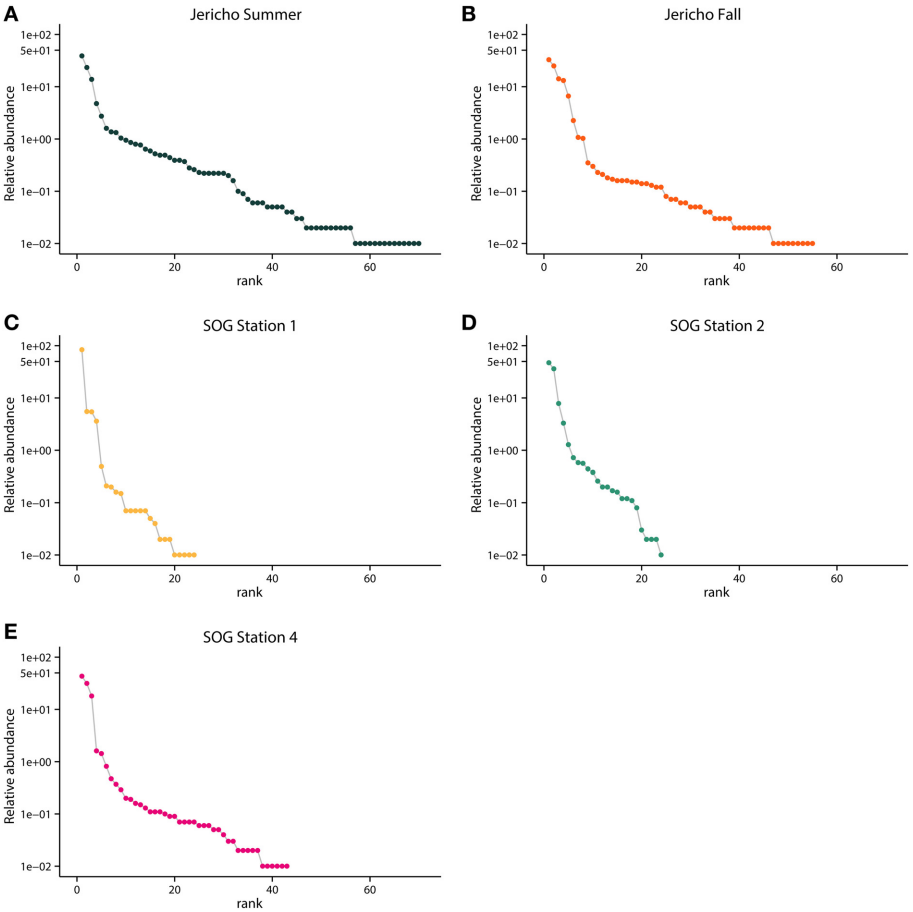


FIGURE 5 | Rank abundance by site. Relative abundance of OTUs in each sample ordered by rank abundance: **(A)** Jericho Summer, **(B)** Jericho Fall, **(C)** SOG Station 1, **(D)** SOG Station 2, **(E)** SOG Station 4. OTUs were clustered at 95% amino acid similarity and OTU relative abundances were normalized to the sample with the lowest number of reads.

type virus of the family *Marnaviridae* (Lang et al., 2004); it has a genome of about 9.1 kb and a high burst size as indicated by the large crystalline arrays of particles in the cytoplasm of infected cells (Tai et al., 2003). HaRNAV was isolated from coastal waters in British Columbia, Canada (Tai et al., 2003) and can remain infectious for many years in sediments (Lawrence and Suttle, 2004). Interestingly, HaRNAV was isolated from the same area as the present study, and appeared ancestral to many of the recovered sequences based on the phylogeny. For example, OTU 0 was most abundant (18,034 reads after rarefaction) and clustered closely with HaRNAV, although the sequence was only 76.4% similar at the amino-acid level. However, other OTUs in the cluster ranged between 55 and 79% similar to HaRNAV, which is low compared to amino-acid similarities of other RNA viruses within a family that usually have greater than 90% aa similarity (Ng et al., 2012).

DISTINCT COMMUNITIES OCCURRED IN DIFFERENT SEASONS AT THE SAME LOCATION

While only 8.6% of the OTUs from Jericho Pier were shared between dates, both samples had similar evenness, although the summer sample had greater richness (**Figure 5**, **Table 2**). The small overlap in OTUs between sampling dates is not surprising given the very different conditions between July and October (**Figure 1**), and the dynamic nature of planktonic communities in response to environmental changes. At the same location, dsDNA viruses belonging to the *Phycodnaviridae*, which infect eukaryotic phytoplankton, varied seasonally based on fingerprint analyses of DNA polymerase gene fragments using denaturing gradient gel electrophoresis; however, some OTUs persisted for extended periods (Short and Suttle, 2003). Similarly, the composition of other aquatic viral communities has been shown to be dynamic although some OTUs persist (Djikeng et al., 2009; Rodriguez-Brito et al., 2010), and in some cases have repeatable seasonal patterns (Chow and Fuhrman, 2012; Clasen et al., 2013; Marston et al., 2013). With only single samples from summer and fall, inferences about dynamics cannot be made from our data.

One of the few taxonomic groups that occurred in both the summer (JP-S) and fall (JP-F) samples from Jericho Pier was related to HaRNAV (**Figure 6**). There was greater diversity of OTUs in this clade in JP-F, even though JP-S had higher richness and higher phylogenetic diversity overall. This is unlike bacterial and phytoplankton communities that tend to be more diverse in winter (Zingone et al., 2009; Ladau et al., 2013). However, the

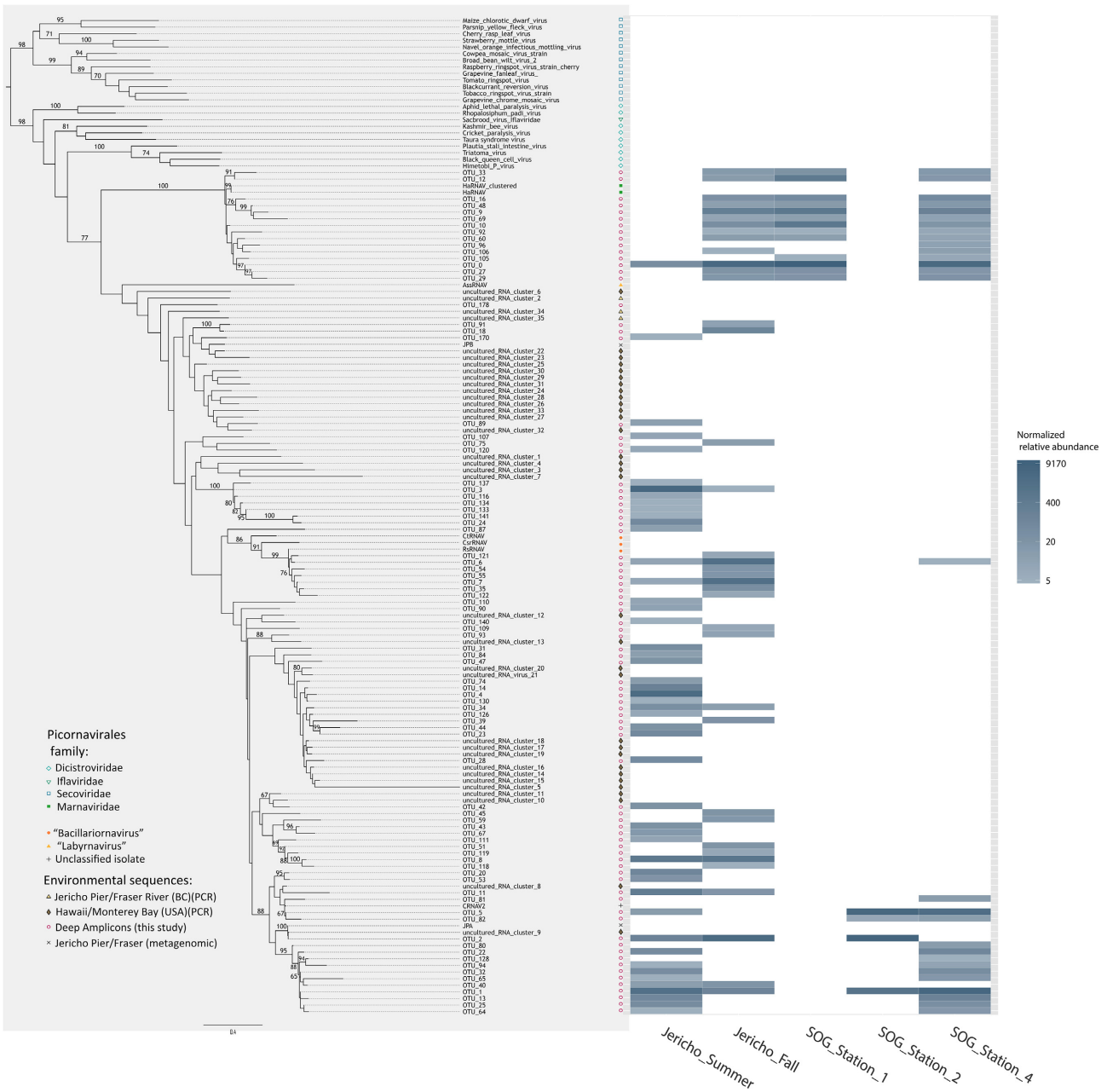


FIGURE 6 | Phylogenetic tree with heatmap. Maximum likelihood (RA×ML) tree of RdRp OTUs that contained more than 5 reads and relevant *Picornavirales* sequences. Bootstrap values above 65% are labeled. Adjacent to the tips of the tree is a heatmap displaying the relative abundance of each OTU at 95% similarity by site. OTU relative abundances were normalized to the sample with the lowest number of reads.

RdRp primers target a specific subset of the viral community that might not reflect the overall taxonomic diversity.

Based on genome organization and sequence identity, RNA viruses that infect diatoms have been assigned to the genus *Bacillarnavirus*, that includes *Rhizosolenia setigera* RNA virus (RsRNAV) (Nagasaki et al., 2004), *Chaetoceros tenuissimus* Meunier RNA virus (CtenRNAV) (Shirai et al., 2008) and *Chaetoceros socialis* f. radians RNA virus (CsfrRNAV) (Tomaru et al., 2009). In the JP-F sample, the most relatively abundant cluster grouped with RsRNAV that infects the marine diatom *Rhizosolenia setigera* (Nagasaki et al., 2004). This corresponded with the highest levels of nitrate + nitrite, which is often associated with high diatom abundances (Zingone et al., 2009); hence, these OTUs are likely associated with viruses infecting diatoms.

DISTINCT COMMUNITIES OCCURRED AT GEOGRAPHICALLY PROXIMATE SITES

Areas of higher habitat diversity, such as stratified water layers, generally have higher biological richness (Klopfer and MacArthur, 1960; Chesson, 2000), and this is consistent with

Table 2 | Phylogenetic diversity, species richness.

	Phylogenetic diversity	Species Richness
Jericho Pier Summer	16.25	46.00
Jericho Pier Fall	13.62	30.00
SOG Station 1	3.89	9.00
SOG Station 2	3.26	4.00
SOG Station 4	8.35	24.00

Phylogenetic diversity (PD) is calculated as in Faith (1992). OTUs were the same as used in the construction of the phylogenetic tree and must have included 5 or more reads.

the much higher richness and phylogenetic diversity found at SOG-4, which was the most stratified site and included the most abundant OTUs from SOG-1 and SOG-2. Most OTUs from SOG-1 clustered in the *Marnaviridae* clade, while most SOG-2 OTUs clustered in a phylogenetically distant clade. Given that we have used very conservative clustering, and that dsDNA viruses infecting phytoplankton are strain specific and have phylogenies that are congruent with their hosts (Clasen and Suttle, 2009; Bellec et al., 2014), and that RNA viruses infecting diatoms and the dinoflagellate, *Heterocapsa circularisquama* (Nagasaki et al., 2005) are host-specific, it implies that closely related OTUs infect closely related taxa of phytoplankton. Hence, it suggests that the most abundant viruses at these three locations infect different species.

There are few clear patterns in the spatial distribution of viruses in marine waters where geographically distant sites are connected by currents and mixing. The best examples are for cyanophages. For instance, when looking at local variation in cyanophages isolated at sites in Southern New England, 72% of the viral OTUs were shared between at least 2 sites (Marston et al., 2013); however, between Bermuda and Southern New England only 2 OTUs overlapped and they comprised only 0.6% of the isolates. Yet, clear patterns of cyanophage OTU distribution by depth occurred in areas adjacent to the SOG when assessed using community fingerprinting (Frederickson et al., 2003). The biggest differences with depth occurred in stratified water in which some OTUs were present at all depths, while others were only present at specific depths, even though the samples were collected only meters apart (Frederickson et al., 2003). These viruses infect cyanobacteria, as opposed to the picorna-like viruses, which likely infect protistan plankton. Nonetheless, the factors governing the distribution of cyanobacterial and protistan hosts are likely similar; hence, different OTUs would be expected to occur in environments with different vertical structure (stratification) of the water column.

Rank abundance curves showed that SOG-1 and SOG-2 were the least even communities (**Figure 5**). Overall, at most sites four to five viral OTUs were most abundant (**Figure 5**) similar to other reports for aquatic viral communities in which a few viruses dominate, but most of the diversity comes from rarer viruses (Angly et al., 2006; Suttle, 2007). Our targeted approach showed that the picornavirus-like virus communities at SOG and JP were dominated by only a few genotypes, supporting previous metagenomic results showing that the OTU distributions of RNA viruses in SOG and JP were highly uneven with little overlap between sites (Culley et al., 2006).

EACH OTU LIKELY REPRESENTS A SINGLE LYTIC EVENT

Given that the hosts of marine *Picornavirales* isolates are protists, and that protists are the most abundant eukaryotes in the sea, it is likely that the majority of OTUs recovered in this study are from viruses that infect these unicellular marine eukaryotes. These eukaryotic communities are highly dynamic and change throughout the year based on environmental and biological factors (Larsen et al., 2004). Since viral infection is usually host specific, the diversity in marine viral communities is a reflection of the underlying diversity of the marine eukaryotic hosts. Moreover, viral propagation is dependent on host encounter rates and is proportional to host-cell abundance (Murray and Jackson, 1992); hence the most abundant taxa will be most likely to encounter and propagate a viral infection, giving the opportunity for rarer species to increase in abundance and promoting diversity (Thingstad, 2000; Winter et al., 2010). Since our study was not over time it is difficult to evaluate whether these data support the Bank model (Breitbart and Rohwer, 2005), however, some taxa were found at one site, but not a similar nearby site, thus these taxa could be present at background levels at some sites and more abundant in others.

It is probable that the most abundant OTUs in these data are from recent lysis of host taxa. An error rate for replication of RNA viruses of about 1 bp mutation per generation (9000 bp genome $\times$ 0.0001 error rate per base pair $=$ 1 bp; Holmes, 2009), and a lower-end burst-size estimate of 1000 particles for marine viruses in the *Picornavirales* that infect protists (Lang et al., 2009), would produce about 1000 different genomes from each lysed cell. For the amplified 500 bp RdRp gene fragment there is a 0.00056% chance of an error in 1 generation, assuming that mutations are distributed evenly in the genome (Sanjuan et al., 2010; Combe and Sanjuán, 2014). Consequently, even with the relatively high error rates of RNA replication, when grouped at 95% similarity at the amino acid level, all of the sequences from a lytic event should fall within a single OTU. The half-life for decay of viral infectivity and particles in the surface mixed layer is typically a few hours (Heldal and Bratbak, 1991; Suttle and Chen, 1992; Noble and Fuhrman, 1997; Bettarel et al., 2009); thus the recovered viral OTUs were likely from recent lytic events. Furthermore, considering the specificity of viruses infecting protists (Short, 2012), each OTU probably stems from viruses infecting a single host taxon. Thus, these data imply that infection of marine protists by viruses in the *Picornavirales* is not only pervasive, but likely involves a wide diversity of host taxa; hence, these viruses are likely important structuring elements for phytoplankton communities that influence nutrient cycling and energy flow.

AMPLICON DEEP SEQUENCING AS AN APPROACH FOR ESTIMATING VIRAL DIVERSITY

Amplicon deep sequencing is a sensitive and high-resolution approach for examining microbial community dynamics over time and space (Caporaso et al., 2011; Gobet et al., 2012; Gibbons et al., 2013). Careful quality trimming of sequences

and removal of singletons is essential for reliable results (Zhou et al., 2011) since errors in sequences will inflate estimates of diversity. With careful data processing and analysis, amplicon deep sequencing is as accurate for assessing community composition and diversity as cloning and Sanger sequencing (Amend et al., 2010), but with much greater depth of coverage of the community.

There are potential biases associated with reverse transcription with random hexamers (which can decrease yield and could inflate diversity) (Zhang and Byrne, 1999), template amplification by PCR (Lee et al., 2012) and with using highly degenerate primers that target a specific part of the community containing many different templates (Culley and Steward, 2007). A danger of the high cycle number can be diversity overestimates which can come from the increasing number of chimeric sequences produced with greater cycle number (Qiu et al., 2001). The sequences were processed with caution considering the high number of PCR cycles employed in this study. Chimera checking *denovo* was used to look for chimeric sequences originating from two higher abundance reads, and reference-based chimera checking was used a database of RdRps from isolated viruses to correct for this potential error. In addition, a conservative cut-off was used of only OTUs comprising more than 5 reads that aligned to the conserved domain alignment.

Read abundance of OTUs can be considered semi-quantitative and good for comparisons of richness and diversity among samples (but not for absolute counts of genes) (Amend et al., 2010; Pinto and Raskin, 2012; Ibarbalz et al., 2014). Moreover, by using control sequences obtained by cloning and Sanger sequencing alongside pyrosequenced libraries containing the same sequence (Supplemental Methods, Figure S2) we verified that amplicon deep sequencing and our sequence processing methodology recovered accurate environmental viral sequences and non-inflated estimates of richness like in studies for bacterial amplicons (Sogin et al., 2006; Huse et al., 2008; Kirchman et al., 2010; Caporaso et al., 2011) and clinical viral studies (Romano et al., 2013; Watson et al., 2013).

CONCLUSION

Amplicon deep sequencing of RdRp gene fragments using 454 pyrosequencing revealed the richness and population structure of marine *Picornavirales* in five coastal samples. The known diversity of viruses in this group was greatly increased with 145 OTUs that differed by at least 5% at the amino-acid level. There were between 24 and 71 OTUs in each sample, with distinct patterns of OTU distribution, richness and diversity among samples. There was little overlap between viral OTUs collected at the same site in summer and fall, and among samples collected 20 km apart on the same day. The high temporal and spatial diversity in RdRp sequences is consistent with viral communities that turnover rapidly, and episodic infection of a wide diversity of protistan hosts. The low overlap in OTUs and phylogenetic diversity among samples implies a dynamic landscape of viral infection and supports the idea that marine picorna-like viruses are important pathogens of marine protists that have an important role in structuring marine planktonic communities, and in nutrient cycling and energy transfer among trophic levels. Ultimately, further study is needed to disentangle the temporal and spatial drivers of these communities.

ACKNOWLEDGMENTS

Thanks to R. Adelshin, A. M. Chan, C. Charlesworth, C. Chénard, R. Cruz, J. Finke, J. M. Labonté, J. Li, T. Nelson, J. P. Payet, E. J. Shelford, M. Vlok, R. A. White III, Q. Zhang, for help with sampling, and sample processing, the crew of the CGCS Vector for facilitating sample collection, and C. Payne for help with sampling, instrumentation and data acquisition, as well as processing of nutrient samples. M. Vlok and R. A. White III contributed to library preparation. Thanks to C. Chénard, C. Chow, J. Finke, T. Heger and E. J. Shelford for insightful comments on the manuscript. Funding was provided by NSERC Discovery and ship-time grants and from the Tula Foundation (Curtis A. Suttle), and NSERC PGS-M and PGS-D scholarships (Julia A. Gustavsen), and a UBC 4-Year Fellowship (Julia A. Gustavsen).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00703/abstract

REFERENCES

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. doi: 10.1016/S0022-2836(05)80360-2

Amend, A. S., Seifert, K. A., and Bruns, T. D. (2010). Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Mol. Ecol.* 19, 5555–5565. doi: 10.1111/j.1365-294X.2010.04898.x

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Bellec, L., Clerissi, C., Edern, R., Foulon, E., Simon, N., Grimsley, N., et al. (2014). Cophylogenetic interactions between marine viruses and eukaryotic picophytoplankton. *BMC Evol. Biol.* 14:59. doi: 10.1186/1471-2148-14-59

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2007). GenBank. *Nucleic Acids Res.* 35, D21–D25. doi: 10.1093/nar/gkl986

Bergh, Ø., Børsheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature* 340, 467–468. doi: 10.1038/340467a0

Bettarel, Y., Bouvier, T., and Bouvy, M. (2009). Viral persistence in water as evaluated from a tropical/temperate cross-incubation. *J. Plankton Res.* 31, 909–916. doi: 10.1093/plankt/fbp041

Bratbak, G., Egge, J., and Heldal, M. (1993). Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination algal blooms. *Mar. Ecol. Prog. Ser.* 93, 39–48. doi: 10.3354/meps093039

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* 13, 278–284. doi: 10.1016/j.tim.2005.04.003

Brussaard, C. P. (2004). Viral control of phytoplankton populations—a review. *J. Eukaryot. Microbiol.* 51, 125–138. doi: 10.1111/j.1550-7408.2004.tb00537.x

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. doi: 10.1016/j.virol.2003.10.033

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. doi: 10.1093/bioinformatics/btp348

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4516–4522. doi: 10.1073/pnas.1000080107

Chen, F., Wang, K., Huang, S., Cai, H., Zhao, M., Jiao, N., et al. (2009). Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environ. Microbiol.* 11, 2884–2892. doi: 10.1111/j.1462-2920.2009.02033.x

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* 31, 343–366. doi: 10.1146/annurev.ecolsys.31.1.343

Chow, C.-E. T., and Fuhrman, J. A. (2012). Seasonality and monthly dynamics of marine myovirus communities: marine myovirus community dynamics at SPOT. *Environ. Microbiol.* 14, 2171–2183. doi: 10.1111/j.1462-2920.2012.02744.x

Clasen, J., Hanson, C., Ibrahim, Y., Weihe, C., Marston, M., and Martiny, J. (2013). Diversity and temporal dynamics of Southern California coastal marine cyanophage isolates. *Aquat. Microb. Ecol.* 69, 17–31. doi: 10.3354/ame01613

Clasen, J. L., and Suttle, C. A. (2009). Identification of freshwater Phycodnaviridae and their potential phytoplankton hosts, using DNA pol sequence fragments and a genetic-distance analysis. *Appl. Environ. Microbiol.* 75, 991–997. doi: 10.1128/AEM.02024-08

Combe, M., and Sanjuán, R. (2014). Variation in RNA virus mutation rates across host cells. *PLoS Pathog.* 10:e1003855. doi: 10.1371/journal.ppat.1003855

Culley, A. I., Lang, A., and Suttle, C. A. (2003). High diversity of unknown picorna-like viruses in the sea. *Nature* 424, 1054–1057. doi: 10.1038/nature01933

Culley, A. I., Lang, A., and Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* 312, 1795–1798. doi: 10.1126/science.1127404

Culley, A. I., Mueller, J. A., Belcaid, M., Wood-Charlson, E. M., Poisson, G., and Steward, G. F. (2014). The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *MBio* 5:e01210–14–e01210–14. doi: 10.1128/mBio.01210-14

Culley, A. I., and Steward, G. F. (2007). New genera of RNA viruses in subtropical seawater, inferred from polymerase gene sequences. *Appl. Environ. Microbiol.* 73, 5937–5944. doi: 10.1128/AEM.01065-07

Darriba, D., Taboada, G. L., Doallo, R., and Bangor, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27, 1164–1165. doi: 10.1093/bioinformatics/btr088

Derelle, E., Ferraz, C., Escande, M.-L., Eycheni,é, 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. doi: 10.1371/journal.pone.0002250

Djikeng, A., Kuzmickas, R., Anderson, N. G., and Spiro, D. J. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS ONE* 4:e7264. doi: 10.1371/journal.pone.0007264

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* 61, 1–10. doi: 10.1016/0006-3207(92)91201-3

Frederickson, C. M., Short, S. M., and Suttle, C. A. (2003). The physical environment affects cyanophage communities in British Columbia inlets. *Microb. Ecol.* 46, 348–357. doi: 10.1007/s00248-003-1010-2

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548. doi: 10.1038/21119

Gibbons, S. M., Caporaso, J. G., Pirrung, M., Field, D., Knight, R., and Gilbert, J. A. (2013). Evidence for a persistent microbial seed bank throughout the global ocean. *Proc. Natl. Acad. Sci. U.S.A.* 110, 4651–4655. doi: 10.1073/pnas.1217767110

Gobet, A., Böer, S. I., Huse, S. M., van Beusekom, J. E., Quince, C., Sogin, M. L., et al. (2012). Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *ISME J.* 6, 542–553. doi: 10.1038/ismej.2011.132

Heldal, M., and Bratbak, G. (1991). Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser.* 72, 205–212. doi: 10.3354/meps072205

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

Huang, S., Wilhelm, S. W., Jiao, N., and Chen, F. (2010). Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences. *ISME J.* 4, 1243–1251. doi: 10.1038/ismej.2010.56

Huse, S. M., Dethlefsen, L., Huber, J. A., Mark Welch, D., Welch, D. M., Relman, D. A., et al. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4:e1000255. doi: 10.1371/journal.pgen.1000255

Ibarbalz, F. M., Pérez, M. V., Figuerola, E. L. M., and Erijman, L. (2014). The bias associated with amplicon sequencing does not affect the quantitative assessment of bacterial community dynamics. *PLoS ONE* 9:e99722. doi: 10.1371/journal.pone.0099722

Jameson, E., Mann, N. H., Joint, I., Sambles, C., and Mühling, M. (2011). The diversity of cyanomyovirus populations along a North–South Atlantic Ocean transect. *ISME J.* 5, 1713–1721. doi: 10.1038/ismej.2011.54

Kemp, P. F., and Aller, J. Y. (2004). Estimating prokaryotic diversity: when are 16 s rDNA libraries large enough? *Limnol. Oceanogr. Methods* 2, 114–125. doi: 10.4319/lom.2004.2.114

Kirchman, D. L., Cottrell, M. T., and Lovejoy, C. (2010). The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ. Microbiol.* 12, 1132–1143. doi: 10.1111/j.1462-2920.2010.02154.x

Klopfer, P. H., and MacArthur, R. H. (1960). Niche size and faunal diversity. *Am. Nat.* 94, 293–300. doi: 10.1086/282130

Koonin, E. V. (1991). The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses. *J. Gen. Virol.* 72, 2197–2206. doi: 10.1099/0022-1317-72-9-2197

Ladau, J., Sharpton, T. J., Finucane, M. M., Jospin, G., Kembel, S. W., O'Dwyer, J., et al. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *ISME J.* 7, 1669–1677. doi: 10.1038/ismej.2013.37

Lang, A. S., Culley, A. I., and Suttle, C. A. (2004). Genome sequence and characterization of a virus (HaRNAV) related to picorna-like viruses that infects the marine toxic bloom-forming alga Heterosigma akashiwo. *Virology* 320, 206–217. doi: 10.1016/j.virol.2003.10.015

Lang, A. S., Rise, M. L., Culley, A. I., and Steward, G. F. (2009). RNA viruses in the sea. *FEMS Microbiol. Rev.* 33, 295–323. doi: 10.1111/j.1574-6976.2008.00132.x

Larsen, A., Flaten, G. A. F., Sandaa, R.-A., Castberg, T., Thyrhaug, R., Erga, S. R., et al. (2004). Spring phytoplankton bloom dynamics in Norwegian coastal waters: microbial community succession and diversity. *Limnol. Oceanogr.* 49, 180–190. doi: 10.4319/lo.2004.49.1.0180

Lawrence, J. E., and Suttle, C. A. (2004). Effect of viral infection on sinking rates of *Heterosigma akashiwo* and its implications for bloom termination. *Aquat. Microb. Ecol.* 37, 1–7. doi: 10.3354/ame037001

Le Gall, O., Christian, P., Fauquet, C. M., King, A. M. Q., Knowles, N. J., Nakashima, N., et al. (2008). *Picornavirales*, a proposed order of positive-sense single-stranded RNA viruses with a pseudo-T = 3 virion architecture. *Arch. Virol.* 153, 715–727. doi: 10.1007/s00705-008-0041-x

Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R., et al. (2012). Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS ONE* 7:e44224. doi: 10.1371/journal.pone.0044224

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., et al. (2011). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Marston, M. F., Taylor, S., Sme, N., Parsons, R. J., Noyes, T. J. E., and Martiny, J. B. H. (2013). Marine cyanophages exhibit local and regional biogeography: biogeography of marine cyanophages. *Environ. Microbiol.* 15, 1452–1463. doi: 10.1111/1462-2920.12062

Mayer, J. A., and Taylor, F. J. R. (1979). A virus which lyses the marine nanoflagellate *Micromonas pusilla*. *Nature* 281, 299–301. doi: 10.1038/281299a0

Mühling, M., Fuller, N. J., Millard, A., Somerfield, P. J., Marie, D., Wilson, W. H., et al. (2005). Genetic diversity of marine *Synechococcus* and co-occurring cyanophage communities: evidence for viral control of phytoplankton. *Environ. Microbiol.* 7, 499–508. doi: 10.1111/j.1462-2920.2005.00713.x

Murray, A. G., and Jackson, G. A. (1992). Viral dynamics: a model of the effects of size, shape, motion and abundance of single-celled planktonic organisms and other particles. *Mar. Ecol. Prog. Ser.* 89, 103–116. doi: 10.3354/meps089103

Nagasaki, K., Ando, M., Itakura, S., Imai, I., and Ishida, Y. (1994). Viral mortality in the final stage of *Heterosigma akashiwo* (Raphidophyceae) red tide. *J. Plankton Res.* 16, 1595–1599. doi: 10.1093/plankt/16.11.1595

Nagasaki, K., Shirai, Y., Takao, Y., Mizumoto, H., Nishida, K., and Tomaru, Y. (2005). Comparison of genome sequences of single-stranded RNA viruses infecting the bivalve-killing dinoflagellate *Heterocapsa circularisquama*. *Appl. Environ. Microbiol.* 71, 8888–8894. doi: 10.1128/AEM.71.12.8888-8894.2005

Nagasaki, K., Tomaru, Y., Katanozaka, N., Shirai, Y., Nishida, K., Itakura, S., et al. (2004). Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. *Appl. Environ. Microbiol.* 70, 704–711. doi: 10.1128/AEM.70.2.704-711.2004

Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., et al. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *J. Virol.* 86, 12161–12175. doi: 10.1128/JVI.00869-12

Noble, R. T., and Fuhrman, J. A. (1997). Virus decay and its causes in coastal waters. *Appl. Environ. Microbiol.* 63, 77–83.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., et al. (2013). *Vegan: Community Ecology Package.* R package version 2.0-10. Available online at: http://CRAN.R-project.org/package=vegan

Paradis, E. (2012). *Analysis of Phylogenetics and Evolution with R*. New York, NY: Springer. doi: 10.1007/978-1-4614-1743-9

Parsons, T. R., Maita, Y., and Lalli, C. M. (1984). *A Manual of Chemical and Biological Methods for Seawater Analysis*. New York, NY: Pergamon Press.

Pinto, A. J., and Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE* 7:e43093. doi: 10.1371/journal.pone.0043093

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Qiu, X., Wu, L., Huang, H., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., et al. (2001). Evaluation of PCR-Generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.* 67, 880–887. doi: 10.1128/AEM.67.2.880-887.2001

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Reeder, J., and Knight, R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods* 7, 668–669. doi: 10.1038/nmeth0910-668b

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J.* 4, 739–751. doi: 10.1038/ismej.2010.1

Romano, C. M., Lauck, M., Salvador, F. S., Lima, C. R., Villas-Boas, L. S., Araújo, E. S. A., et al. (2013). Inter- and intra-host viral diversity in a large seasonal DENV2 outbreak. *PLoS ONE* 8:e70318. doi: 10.1371/journal.pone.0070318

Sanfaçon, H., Wellink, J., Gall, O., Karasev, A., Vlugt, R., and Wetzel, T. (2009). Secoviridae: a proposed family of plant viruses within the order *Picornavirales* that combines the families *Sequiviridae* and *Comoviridae*, the unassigned genera *Cheravirus* and *Sadwavirus*, and the proposed genus Torradovirus. *Arch. Virol.* 154, 899–907. doi: 10.1007/s00705-009-0367-z

Sanjuan, R., Nebot, M. R., Chirico, N., Mansky, L. M., and Belshaw, R. (2010). Viral mutation rates. *J. Virol.* 84, 9733–9748. doi: 10.1128/JVI.00694-10

Schroeder, D. C., Oke, J., Hall, M., Malin, G., and Wilson, W. H. (2003). Virus succession observed during an *Emiliania huxleyi* bloom. *Appl. Environ. Microbiol.* 69, 2484–2490. doi: 10.1128/AEM.69.5.2484-2490.2003

Shirai, Y., Tomaru, Y., Takao, Y., Suzuki, H., Nagumo, T., and Nagasaki, K. (2008). Isolation and characterization of a single-stranded RNA virus infecting the marine planktonic diatom *Chaetoceros tenuissimus* Meunier. *Appl. Environ. Microbiol.* 74, 4022–4027. doi: 10.1128/AEM.00509-08

Short, S. M. (2012). The ecology of viruses that infect eukaryotic algae: Algal virus ecology. *Environ. Microbiol.* 14, 2253–2271. doi: 10.1111/j.1462-2920.2012.02706.x

Short, S. M., and Suttle, C. A. (2003). Temporal dynamics of natural communities of marine algal viruses and eukaryotes. *Aquat. Microb. Ecol.* 32, 107–119. doi: 10.3354/ame032107

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Syst. Biol.* 57, 758–771. doi: 10.1080/10635150802429642

Steward, G. F., Culley, A. I., Mueller, J. A., Wood-Charlson, E. M., Belcaid, M., and Poisson, G. (2012). Are we missing half of the viruses in the ocean? *ISME J.* 7, 672–679. doi: 10.1038/ismej.2012.121

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

Suttle, C. A. (2005). Viruses in the sea. *Nature* 437, 356–361. doi: 10.1038/nature04160

Suttle, C. A., Chan, A. M., and Cottrell, M. T. (1991). Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl. Environ. Microbiol.* 57, 721–726.

Suttle, C. A., and Chen, F. (1992). Mechanisms and rates of decay of marine viruses in seawater. *Appl. Environ. Microbiol.* 58, 3721–3729.

Tai, V., Lawrence, J. E., Lang, A. S., Chan, A. M., Culley, A. I., and Suttle, C. A. (2003). Characterization of HaRNAV, a single-stranded RNA virus causing lysis of *Heterosigma akashiwo* (Raphidophyceae). *J. Phycol.* 39, 343–352. doi: 10.1046/j.1529-8817.2003.01162.x

Takao, Y., Mise, K., Nagasaki, K., Okuno, T., and Honda, D. (2006). Complete nucleotide sequence and genome organization of a single-stranded RNA virus infecting the marine fungoid protist *Schizochytrium* sp. *J. Gen. Virol.* 87, 723–733. doi: 10.1099/vir.0.81204-0

Thingstad, T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* 45, 1320–1328. doi: 10.4319/lo.2000.45.6.1320

Tomaru, Y., Katanozaka, N., Nishida, K., Shirai, Y., Tarutani, K., Yamaguchi, M., et al. (2004). Isolation and characterization of two distinct types of HcRNAV, a single-stranded RNA virus infecting the bivalve-killing microalga *Heterocapsa circularisquama*. *Aquat. Microb. Ecol.* 34, 207–218. doi: 10.3354/ame034207

Tomaru, Y., Takao, Y., Suzuki, H., Nagumo, T., and Nagasaki, K. (2009). Isolation and characterization of a single-stranded RNA virus infecting the bloom-forming diatom *Chaetoceros socialis*. *Appl. Environ. Microbiol.* 75, 2375–2381. doi: 10.1128/AEM.02580-08

Van Etten, J. L., Meints, R. H., Burbank, D. E., Kuczmarski, D., Cuppels, D. A., and Lane, L. C. (1981). Isolation and characterization of a virus from the intracellular green alga symbiotic with *Hydra viridis*. *Virology* 113, 704–711. doi: 10.1016/0042-6822(81)90199-9

Wang, K., Wommack, K. E., and Chen, F. (2011). Abundance and distribution of *Synechococcus* spp. and cyanophages in the Chesapeake Bay. *Appl. Environ. Microbiol.* 77, 7459–7468. doi: 10.1128/AEM.00267-11

Watson, S. J., Welkers, M. R. A., Depledge, D. P., Coulter, E., Breuer, J. M., de Jong, M. D., et al. (2013). Viral population analysis and minority-variant detection using short read next-generation sequencing. *Philos. Trans. R. Soc. B Biol. Sci.* 368:20120205. doi: 10.1098/rstb.2012.0205

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer.

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience* 49, 781–788. doi: 10.2307/1313569

Wilkinson, L. (2011). *Venneuler: Venn and Euler Diagrams*. Available online at: http://CRAN.R-project.org/package=venneuler

Winter, C., Bouvier, T., Weinbauer, M. G., and Thingstad, T. F. (2010). Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol. Mol. Biol. Rev.* 74, 42–57. doi: 10.1128/MMBR.00034-09

Yokoyama, R., and Honda, D. (2007). Taxonomic rearrangement of the genus Schizochytrium sensu lato based on morphology, chemotaxonomic characteristics, and 18S rRNA gene phylogeny (Thraustochytriaceae, Labyrinthulomycetes): emendation for Schizochytrium and erection of Aurantiochytrium and Oblongichytrium gen. nov. *Mycoscience* 48, 199–211. doi: 10.1007/S10267-006-0362-0

Zhang, J., and Byrne, C. (1999). Differential priming of RNA templates during cDNA synthesis markedly affects both accuracy and reproducibility of quantitative competitive reverse-transcriptase PCR. *Biochem. J.* 337, 231–241. doi: 10.1042/0264-6021:3370231

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., et al. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J.* 5, 1303–1313. doi: 10.1038/ismej.2011.11

Zingone, A., Dubroca, L., Iudicone, D., Margiotta, F., Corato, F., Ribera d'Alcalà, M., et al. (2009). Coastal phytoplankton do not rest in winter. *Estuaries Coasts* 33, 342–361. doi: 10.1007/s12237-009-9157-9

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

Received: 06 October 2014; accepted: 26 November 2014; published online: 15 December 2014.

*Citation: Gustavsen JA, Winget DM, Tian X and Suttle CA (2014) High temporal and spatial diversity in marine RNA viruses implies that they have an important role in mortality and structuring plankton communities. Front. Microbiol. **5**:703. doi: 10.3389/fmicb.2014.00703*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 22 April 2015
doi: 10.3389/fmicb.2015.00315

Previously unknown evolutionary groups dominate the ssDNA gokushoviruses in oxic and anoxic waters of a coastal marine environment

Jessica M. Labonté[1†], *Steven J. Hallam*[1,2,3] *and Curtis A. Suttle*[1,2,4,5*]

[1] *Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada,* [2] *Canadian Institute for Advanced Research, Toronto, ON, Canada,* [3] *Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada,* [4] *Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada,* [5] *Department of Botany, University of British Columbia, Vancouver, BC, Canada*

OPEN ACCESS

Edited by:
Katrine L. Whiteson, University of California, Irvine, USA

Reviewed by:
Andrew David Millard, University of Warwick, UK
Karyna Rosario, University of South Florida, USA

***Correspondence:**
Curtis A. Suttle, Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia, 2178-2207 Main Mall, Vancouver, BC V6T 1Z4, Canada
suttle@science.ubc.ca

†Present address:
Jessica M. Labonté, Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *27 October 2014*
Accepted: *29 March 2015*
Published: *22 April 2015*

Citation:
Labonté JM, Hallam SJ and Suttle CA (2015) Previously unknown evolutionary groups dominate the ssDNA gokushoviruses in oxic and anoxic waters of a coastal marine environment. Front. Microbiol. 6:315. doi: 10.3389/fmicb.2015.00315

Metagenomic studies have revealed that ssDNA phages from the family *Microviridae* subfamily *Gokushovirinae* are widespread in aquatic ecosystems. It is hypothesized that gokushoviruses occupy specialized niches, resulting in differences among genotypes traversing water column gradients. Here, we use degenerate primers that amplify a fragment of the gene encoding the major capsid protein to examine the diversity of gokushoviruses in Saanich Inlet (SI), a seasonally anoxic fjord on the coast of Vancouver Island, BC, Canada. Amplicon sequencing of samples from the mixed oxic surface (10 m) and deeper anoxic (200 m) layers indicated a diverse assemblage of gokushoviruses, with greater richness at 10 m than 200 m. A comparison of amplicon sequences with sequences selected on the basis of RFLP patterns from eight surface samples collected over a 1-year period revealed that gokushovirus diversity was higher in spring and summer during stratification and lower in fall and winter after deep-water renewal, consistent with seasonal variability within gokushovirus populations. Our results provide persuasive evidence that, while specific gokushovirus genotypes may have a narrow host range, hosts for gokushoviruses in SI consist of a wide range of bacterial taxa. Indeed, phylogenetic analysis of clustered amplicons revealed at least five new phylogenetic groups of previously unknown sequences, with the most abundant group associated with viruses infecting SUP05, a ubiquitous and abundant member of marine oxygen minimum zones. Relatives of SUP05 dominate the anoxic SI waters where they drive coupled carbon, nitrogen, and sulfur transformations along the redoxline; thus, gokushoviruses are likely important mortality agents of these bacteria with concomittant influences on biogeochemical cycling in marine oxygen minimum zones.

Keywords: *Gokushovirinae*, host range, viral diversity, SUP05, oxygen minimum zones, PCR sequencing

Introduction

Bacteriophages belonging to the *Microviridae* family consist of a ∼30 nm icosahedral capsid containing a positive-sense ssDNA molecule of 4.5–6.1 kb (King et al., 2012). Replication requires a minimum of two coat proteins (VP1 and VP2), a scaffolding protein (VP3), a replication protein (VP4), and a DNA packaging protein (VP5). Based on the phylogeny of the major capsid protein (MCP or VP1) from isolates, the *Microviridae* family is divided into two groups (Brentlinger et al., 2002); the *Microvirus* genus contains phages like phiX174 and G4 that infect *Escherichia coli* (Godson et al., 1978), while the *Gokushovirinae* subfamily includes those infecting parasitic bacteria such as *Chlamydia* [Chp1 (Storey et al., 1989), Chp2 (Liu et al., 2000; Everson et al., 2002), Chp3 (Garner et al., 2004)], *Bdellovibrio* [(phiMH2K; Brentlinger et al., 2002)], and *Spiroplasma* [(SpV4; Chipman et al., 1998)]. While it is commonly thought that *Microviridae* phages are strictly lytic (Liu et al., 2000; Garner et al., 2004; Salim et al., 2008), an *in silico* study found sequences with similar genome organization to gokushoviruses associated with *Bacteroidetes* from the human gut and mouth, which suggests that these phages can be temperate (Krupovic and Forterre, 2011). The temperate *Microviridae* phages are phylogenetically distinct from gokushoviruses, and have been assigned to the *Alpavirinae*, a proposed new sub-family within the *Microviridae* (Krupovic and Forterre, 2011). Recently, 81 microvirus genomes (including 42 gokushoviruses) were assembled from various environmental metegenomic data, identifying a new group, the *Pichovirinae*, which harbored a different genome organization of the conserved genes, indicating that microviruses display great diversity and may play an important role in many ecosystems (Roux et al., 2012b).

Marine gokushoviruses were first revealed in viral metagenomic data from the Strait of Georgia (SOG), Gulf of Mexico (GOM), and Sargasso Sea (SAR; Angly et al., 2006), and are among the most commonly recovered sequences from ssDNA phages in marine metagenomic data (Rosario and Breitbart, 2011). They were particularly abundant in the SAR, where 6% of the sequences were similar to the phage Chp1 that infects *Chlamydia psittaci* (Angly et al., 2006). This abundance of ssDNA sequences allowed for the assembly of two environmental *Gokushovirinae* genomes, with the help of PCR amplification (Tucker et al., 2011). A survey in the Atlantic Ocean showed a different depth distribution of these two genomes, consistent with alternative host-infection patterns. Sequences belonging to gokushoviruses have also been found in marine (Labonté and Suttle, 2013b; McDaniel et al., 2014) and fresh waters (López-Bueno et al., 2009; Roux et al., 2012a), stromatolites (Desnues et al., 2008), confined aquifers (Smith et al., 2013), and pelagic sediments (Yoshida et al., 2013). Based on these observations, degenerate primers designed to amplify fragments of the genes encoding the replication initiator (ORF4 or Rep; Tucker et al., 2011) and MCP (Labonté and Suttle, 2013a; Hopkins et al., 2014) were used to examine the distribution and diversity of gokushoviruses in marine ecosystems. Phylogenetic analyses of the amplified fragments revealed that ssDNA phages have different geographic distributions (Labonté and Suttle, 2013a), and that the genetic distance of gokushovirus sequences increased with geographic distance (Tucker et al., 2011). Most of the hosts of environmental gokushoviruses are unknown, but it is hypothesized that they occupy specialized niches, and that specific gokushovirus genotypes have limited geographic range (Angly et al., 2006; Tucker et al., 2011; Labonté and Suttle, 2013a).

Saanich Inlet (SI) is a steep-sided fjord with restricted circulation due to a shallow glacial sill located at the entrance. During spring and summer, high primary productivity in surface waters combined with limited basin circulation contribute to the formation of deep-water anoxia (Anderson, 1973). The anoxic zone is characterized by accumulation of CH_4, NH_3, and H_2S (Anderson, 1973; Lilley et al., 1982; Ward et al., 1989). Typically, in late summer oxygenated nutrient-rich water from Haro Strait (connecting SI to the SOG) cascades over the sill, mixing the oxic, and anoxic waters from top to bottom (Anderson, 1973). Recently, single-cell amplified genomic data (SAGs) from uncultured SUP05 bacteria from marine oxygen minimum zones revealed identical *Microviridae* sequences in 8 of 127 SAGs, suggesting a recent infection event (Roux et al., 2014).

Here, rather than looking at seasonal changes (Labonté and Suttle, 2013a), the diversity of gokushoviruses was examined in the mixed oxic surface (10 m) and deeper anoxic (200 m) layers of SI to better understanding their dynamics and roles in environments with contrasting levels of oxygen. We used degenerate primers for the MCP in combination with amplicon sequencing using 454 technology to reveal a diverse assemblage of gokushoviruses with greater richness at 10 m than 200 m. The results provide persuasive evidence that gokushoviruses likely infect a wide range of hosts, and may be important mortality agents of the SUP05 clade of gamma proteobacteria, an important taxonomic group involved in carbon, nitrogen, and sulfur cycling in marine oxygen minimum zones.

Materials and Methods

Preparation of Marine Samples

On a monthly basis, ∼20 L of water from Station S3 in SI (**Figure 1**) were filtered to remove cells using 0.22-μm pore-size Sterivex^{TM} filter units (Millipore). The viruses were concentrated from the filtrate by tangential flow filtration using a TFF 30-kDa cartridge (Millipore) to a final volume of ∼250 mL, and stored at 4°C until used following the procedure outlined in (Suttle et al., 1991). Amplicons were sequenced from two composite samples of virus concentrates comprised of samples collected from 10 m and 200 m, respectively, during April 2007, and February, March, April, June, August, and November (200 m), or December (10 m) 2008.

ssDNA Purification, PCR Amplification, and Sequencing

For each virus concentrate, viral ssDNA was extracted using a QIAprep Spin M13 kit (Qiagen), according to the manufacturer's protocol. To create double-stranded DNA, 5 μL of the ssDNA

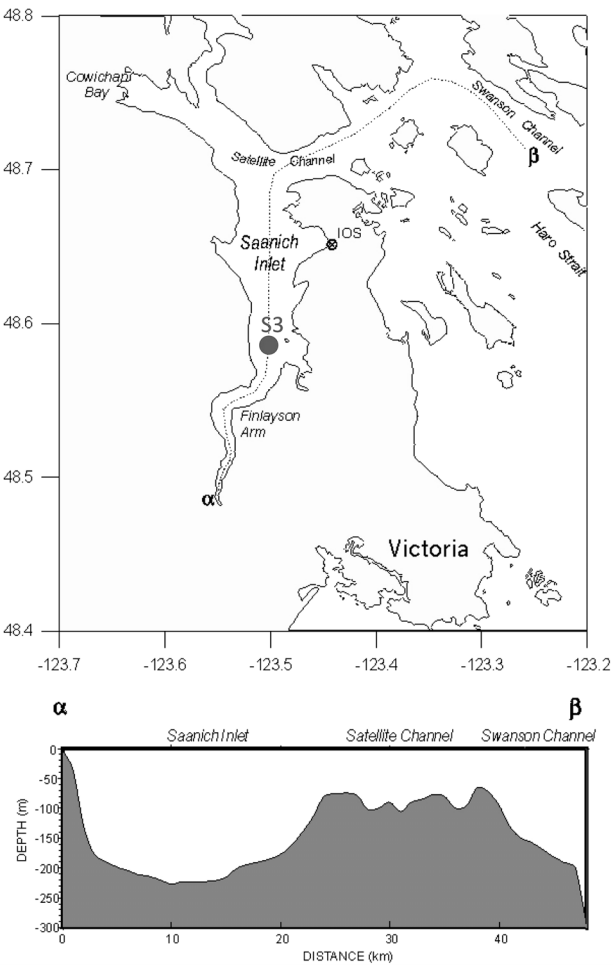


FIGURE 1 | Map of Saanich Inlet (SI; top) and cross-sectional depth profile (bottom). Station S3 (48°35′N, 123°30′W) was sampled in the current study. Map adapted from Fisheries and Oceans Canada (http://www.pac.dfo-mpo.gc.ca/science/oceans/BCinlets/saanich-eng.htm).

preparation was subjected to multiple displacement amplification (MDA; Repli-g Mini kit), and purified using a QIAamp DNA Mini kit. Denaturation of dsDNA was limited during MDA by adding the stop solution N1 immediately after the denaturation solution D1. The purified MDA DNA was resuspended in 100 μL of TE, and 10 μL was used as template in each PCR reaction mixture consisting of Taq DNA polymerase assay buffer [20 mM Tris·HCl (pH 8.4), 50 mM KCl], 1.5 mM MgCl2, 125 μM of each deoxyribonucleoside triphosphate, 1 μM of each primer amplifying a fragment of the MCP gene [MicroVP1–F1 (CGN GCN TAY AAY TTR ATH), VP1–F2 (AGN GCN TAY AAY TTR CTN), MicroVP1–R1 (NCG YTC YTG RTA NCC RAA), and Micro-VP1–R2 (NCT YTC YTG RTA NCC RAA)] and 2.5 U of Platinum® Taq DNA polymerase (Invitrogen). Forward and reverse primer pairs were used to reduce the degeneracy within a single primer. Negative controls contained all reagents except DNA template. The samples were denatured at 94°C for 3 min, followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 50°C for 30 s, and elongation at 72°C for 50 s, with a final elongation step of 72°C for 5 min. The expected PCR product was ∼800 bp in length. PCR amplicons were purified with a MinElute PCR purification kit (Qiagen), pooled into mixes from the 10 or 200 m depth and concentrated using a Millipore YM-30 Microcon centrifugal filter to a final volume of ∼50 μL; a total of 500 ng of DNA amplicons from each pooled 10 and 200 m sample were sent for pyrosequencing using Roche 454 FLX instrumentation with Titanium chemistry at the Broad Institute at the Massachusetts Institute of Technology.

Sequence Binning and Clustering

The sequences were screened for quality and length. Reads were removed if they contained one or more ambiguous bases (Ns), were shorter than 200 nucleotides, or did not match the priming site at the proximal end. Reads were binned based on the primer sequence, which was subsequently trimmed. Sequences arising from both forward primers (F) and both reverse primers (R) were combined resulting in F and R bins. Sequence errors can occur throughout the workflow, including an error rate of 1 per 10^6–10^7 bp (Dean et al., 2001) for WGA, and homopolymers, insertions and deletions of about 0.1% per base for pyrosequencing (Margulies et al., 2005; Huse et al., 2010; Quince et al., 2011); however, the impact of these errors were minimized by clustering the reads into operational taxonomic units (OTUs) at 95% identity using CD-hit (Li and Godzik, 2006). Clustering at 95% also recruited most singletons into an OTU and allowed the data to be compared with the OTUs from a previous seasonal study of gokushoviruses in SI (Labonté and Suttle, 2013a).

Operational taxonomic units were queried in a BLAST search analysis (NCBI BLAST 2.2.2) using an *e*-value cut-off of 10^{-5} against a manually curated database derived from environmental sequences and sequences from gokushovirus isolates composed of all the *Microviridae* genomes available in GenBank (as of November 23, 2014), assembled genomes from Roux et al. (2012b), environmental MCP amplicon sequences from Labonté and Suttle (2013a) and Hopkins et al. (2014). OTUs that did not have a significant hit to sequenced gokushoviruses were removed from the phylogenetic analysis.

Diversity and Species Richness Calculations

For each primer bin, a rank-abundance distribution of phylotypes was generated and subsequently fitted to a power-law function using non-linear regression. For each primer bin, rarefaction species richness curves, and diversity indices were calculated using the Vegan Ecological Diversity package in R (R Development Core Team, 2011). Total estimated richness (S_p) was calculated following Chao's equation (Chao, 1987). The Shannon–Weaver (H') diversity index was calculated as in Hill (1973) on a subsample of 700 reads.

Phylogenetic Analyses

Nucleotide OTUs were aligned with other environmental sequences (Roux et al., 2012b; Yoshida et al., 2013; Hopkins et al., 2014) using MAFFT (Katoh et al., 2002) with the E-INS-I

parameters. We worked with the nucleotide sequences because of problems such as homopolymers associated with the 454 platform made it difficult to accurately infer the correct amino-acid sequences. The alignment of the end product (R primers) was trimmed to get the conserved regions only and phylogenetic analysis was performed with phyML (Guindon et al., 2010) under the HKY85 substitution model with an invgamma distribution with approximate likelihood ratio test (aLRT). Trees were viewed in FigTree[1].

Nucleotide Sequence Accession

Raw sequences, OTU sequences, alignments, and trees are publicly available on Dryad[2].

Results and Discussion

The results from our analyses showed that the oxic and anoxic waters of SI are home to diverse gokushoviruses that comprise at least five previously unknown phylogenetic groups, composed of many numerically dominant OTUs. The most abundant group is associated with viruses that infect SUP05, a group of sulfur-oxidizing bacteria that are ubiquitous and abundant players in marine oxygen minimum zones. These results and their interpretations are detailed below.

Amplicon Sequencing of the MCP From Gokushoviruses

Amplicon sequences of the MCP from two pooled mixes (10 and 200 m) generated 7195 (F) and 2135 (R) good reads (no N, exact primer match, no chimeras) for the 10-m bin and 2687 (F) and 710 (R) good reads for the 200-m bin (**Table 1**). Quality controls using gel electrophoresis and DNA quantification, indicated that the yields from PCR amplification of the 200-m samples were consistently lower than for the 10-m samples (data not shown), which can explain the lower number of MCP amplicons obtained for the 200-m samples.

Reads longer than 200 bp from each bin were clustered into OTUs with more than 95% sequence similarity, and ranked to show the relative abundance of each gokushovirus taxon in our dataset (**Figure 2**). The rank-abundance plots display a similar trend for each primer bin, with a few dominant genotypes and a

[1]http://tree.bio.ed.ac.uk/software/figtree/

[2]http://dx.doi.org/10.5061/dryad.7gg25

TABLE 1 | Number of reads and OTUs after clustering at 95% similarity and OTUs sharing similarity with known *Microviridae* phages.

	Primer	Number of good reads	Number of OTUs	Microviridae OTUs	Most abundant % reads	Singletons % reads	Doublets % reads
10 m	F	7195	504	446	17.9	2.9	2.1
	R	2135	228	213	16.2	4.3	3
200 m	F	2687	117	80	14.8	1.6	0.8
	R	710	36	25	21.5	1.7	1.1

Clustering was performed on sequences that were >200 bp. The number of reads represents the number of reads without ambiguous base.

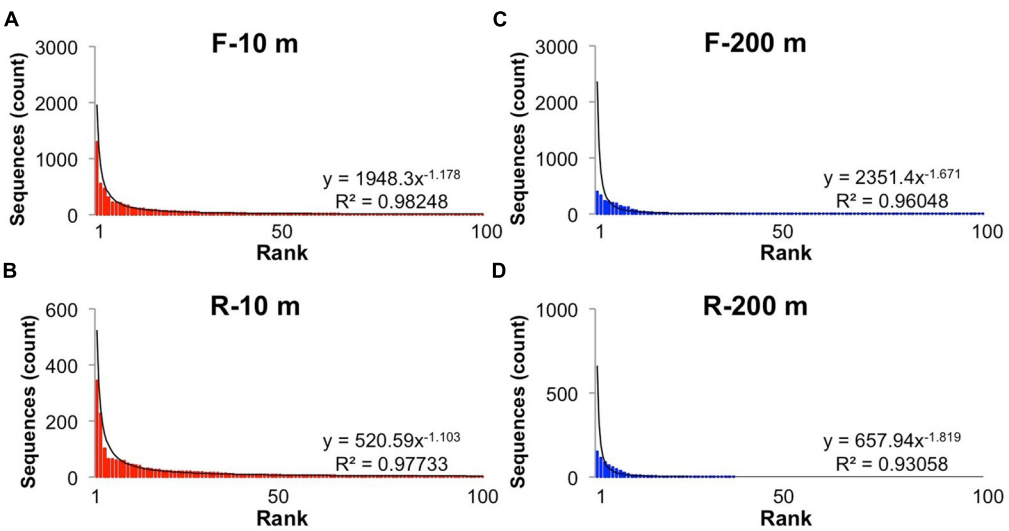


FIGURE 2 | Rank-abundance distribution of operational taxonomic units (OTUs). Rank abundance distribution of the most abundant (up to 100) OTUs (95% similarity) for the forward and reverse primers and each sample: **(A)** Forward primers, 10 m; **(B)** Reverse primers, 10 m, **(C)** Forward primers, 200 m; **(D)** Reverse primers, 200 m. The solid line shows the power-law curve fitting the rank-abundance distribution of phylotypes. Equations and adjusted R^2 are also shown for each curve.

long tail of doubletons and singletons. The rank abundance distribution of genotypes was approximated by a power-law function with R^2-values >0.95. In contrast, there were fewer OTUs in the 200-m samples, and the distribution was less well described by a power-law function. Phages and their hosts usually follow a power-law rank-abundance distribution (Edwards and Rohwer, 2005; Hoffmann et al., 2007; Suttle, 2007). Possible explanations for environmental phage genotypes following a power-law distribution (Edwards and Rohwer, 2005) are that multiple viruses compete for the same hosts or that each virus is specific for one host but the hosts compete for resources. In the first scenario, the most abundant viruses are the most successful at finding and infecting their hosts; hence, every round of infection produces more of that virus. In the second situation, the host that gets more nutrients divides more rapidly resulting in more available hosts for the viruses. In this case, the most abundant viruses are the ones that infect the most abundant bacteria.

The overall sequence similarity within primer bins was compared by clustering the sequences using multiple similarity thresholds (**Figure 3**). Both F and R sequences were more similar to each other at 200 m than at 10 m. This was particularly true for the R sequences (3′ end) for which a decrease in the similarity threshold from 95 to 90% resulted in 41.7% fewer OTUs at 200 m, indicating that many of the sequences were similar and suggesting recent infection events. The 3′ end of the PCR product is more conserved, which can explain the higher similarity among OTUs in the reverse primer bins. For the other samples, the decrease in OTUs was steeper when the similarity threshold was below 80%, indicating the sequences were more diverse.

Rarefaction curves did not plateau (**Figure 4**), indicating that the sequencing depth was inadequate to capture the entire gokushovirus community. Based on the number and frequency of each OTU, the total richness was estimated to be between 355 and 504 OTUs at 10 m, and between 49 and 189 OTUs at 200 m

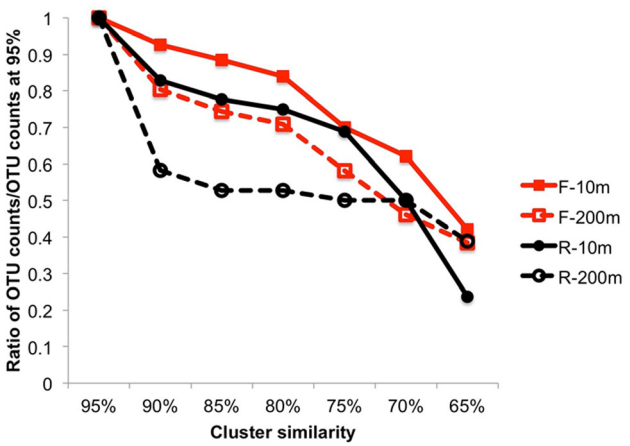


FIGURE 3 | Cluster analysis from 95 to 65% similarity showing the higher similarity within the end of the amplicon sequences and a lower similarity within the beginning of the amplicon sequences. The ratio is expressed as the number of OTUs/number of OTUs at 95%.

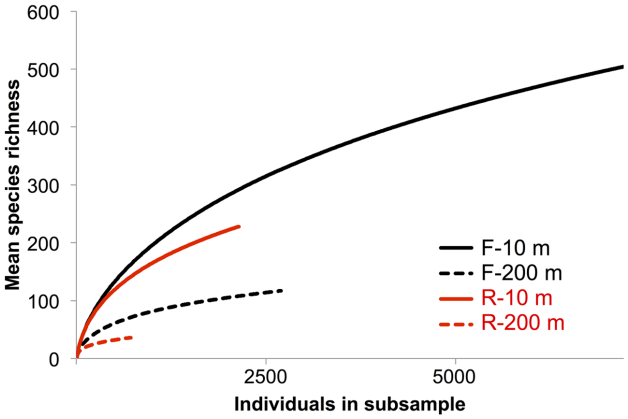


FIGURE 4 | Rarefaction species richness curves. Rarefaction curve analysis to infer the number of observed virus phylotypes present within each gene-specific primer bin. Each curve denotes the number of observed OTUs as a function of the number of sampled reads. Reads were clustered into OTUs based on a 95% similarity threshold.

(**Table 2**). Therefore, between 65 and 62% of total gokushovirus amplicon richness in SI (**Table 2**) was estimated to be captured using the F primers at 10 and 200 m, respectively. Since most reads were not singletons (e.g., only 2.9% of singletons for the F primers), the richness that was not captured likely comprised rare genotypes. Usually, environmental samples are dominated by relatively few genotypes and many more low-abundance ones that account for most of the genetic diversity (Hoffmann et al., 2007; Suttle, 2007). Moreover, abundant and rare genotypes can be temporally and spatially dynamic with a rare genotype being dominant in a different environment or when conditions change (Sogin et al., 2006; Huse et al., 2010; Gobet et al., 2012; Shade et al., 2014).

Care must be taken in interpreting diversity estimates calculated from the data. First, evenness can be influenced by the use of MDA, which can unevenly amplify the initial template (Dean et al., 2001), be biased toward small circular ssDNA molecules and form chimeras (Binga et al., 2008; Polson et al., 2011). Second, the PCR products were pooled from multiple months, which could have affected the evenness observed. For example, if one time point was dominated by a single genotype, it may be highly represented in the pooled sample, even if it was in relatively low abundance at other times. In contrast, a sample with many genotypes at similar abundance will have a lower

TABLE 2 | Observed and estimated richness and diversity indices.

	Primer	Observed richness	Estimated richness (S_{chao})	Shannon diversity index
10 m	F	504	786±54	3.93
	R	228	355±36	3.87
200 m	F	117	189±31	3.03
	R	36	49±10	2.50

The richness is expressed in numbers of OTUs defined at 95% identity.

concentration of each genotype that will be further diluted in the mix. Despite these caveats, and even though temporal variation is integrated across samples, it is reasonable to assume that the observed differences in diversity estimates between the pooled 10 m and 200 m samples is valid. Moreover, because temporal variation is integrated across dates, differences in overall richness are unlikely to be the result of among sample variation; hence, the 4–10-fold higher richness observed in the 10-m samples relative to the 200-m samples (**Table 2**) is likely real. Consistent with this observation, Shannon indices are also higher for the 10-m samples (4.21 and 4.05) than for the 200-m samples (3.18 and 2.54; **Table 2**), reinforcing observations of lower prokaryotic diversity in the anoxic versus oxic waters of SI, with a Shannon index of 0.87 vs. 1.15 for archaea and 0.39 vs. 3.15 at 215 m and 10 m, respectively, (Zaikova et al., 2010).

Seasonal Variation of Gokushoviruses in the Saanich Inlet

In a previous study Labonté and Suttle (2013a) used RFLP analysis to select MCP amplicons for sequencing from SI, SOG, and the GOM. For SI, RFLP analysis of 180 clones from PCR products spanning nine samples resolved 19 unique bands. More bands were observed in spring and summer, when the bacterial abundance is higher and the water column becomes increasingly stratified, while fewer bands were observed in the fall and winter, when bacterial abundance is lower after deep water renewal. Amplicon 454 sequencing recovered 15 of the 19 sequences from SI associated with the RFLPs (**Table 3**). All four sequences associated with a specific RFLP that were not recovered in the 454 data were found only once, and consequently may have been absent from the 454 data. Since both methods used different DNA preparations, the absence of these sequences may have resulted from MDA or PCR biases. It is also possible that the 454 sequencing was not deep enough, as alluded to in the rarefaction curves that did not plateau (**Figure 4**). A lack of sequencing depth could also explain why most of the RFLP sequences were only found with either the F or R primer, but not both. Nonetheless, the richness recovered was much higher using 454 sequencing than by RFLP analysis.

In contrast to the SI results, none of the 13 sequences from the GOM and only two out of 12 sequences from the SOG (SOG3-31 and SOG4-29) that were associated with RFLPs, were recovered in the 454-sequencing data. The SOG sample comprised a mixture of 85 virus concentrates from the SOG and surrounding inlets, including SI. The fact that only two sequences were recovered in the 454 data suggests that gokushovirus sequences display a high degree of endemicity. In contrast, studies on the portal protein from myoviruses (Short and Suttle, 2005; Sullivan et al., 2008), and DNA polymerase B from podoviruses (Breitbart et al., 2004; Chen et al., 2009; Labonté et al., 2009; Huang et al., 2010) and phycodnaviruses (Short and Suttle, 2002) have recovered identical or nearly identical sequences from very different environments.

TABLE 3 | Recovery of RFLP Saanich sequences within the amplicon deep-sequencing database.

RFLP sequence	Apr 08	Jan 09	Mar 09	May 09	Jul 09	Aug 09	Nov 09	OTU	Frequency of OTU	Frequency (%)	Sample
SI-01											
SI-02								DMER9	1/7195	0.01	F-10 m
SI-03											
SI-04											
SI-05								DBSZ3	1/7195	0.01	F-10 m
SI-06								D5JJE	2/2135	0.09	R-10 m
SI-07								ENNJ4	60/2135	2.81	R-10 m
SI-07								DCFCF	1/2135	0.05	F-10 m
SI-08/SI-09								DV3J0	1/7195	0.01	F-10 m
SI-10								D0YTA	20/7195	0.28	F-10 m
SI-11								EW8A2	57/2135	2.67	R-10 m
SI-12								C3HEQ	118/7195	1.64	F-10 m
SI-13	120							DNX08	64/2135	3.00	R-10 m
SI-13	120							F9ET8	49/2687	1.82	F-200 m
SI-14								BOBYV	1/7195	0.01	F1-10 m
SI-16								D0O8I	1/7195	0.01	F1-10 m
SI-17	120										
SI-18								EKHXK	5/7195	0.07	F-10 m
SI-18								D3RZ4	6/2135	0.28	R-10 m
SI-SOG-19								C4GME	3/2135	0.14	R-10 m
SI-SOG-19								DSP87	4/7195	0.06	F-10 m

A total of 15 out of 19 sequences (shown in gray) were sequenced again with amplicon deep sequencing. The seasonal variability of each of these OTUs is represented by dark boxes in the sample months from which they were sequenced. SI, Saanich Inlet; SOG, Strait of Georgia. SI-08 and SI-09 are combined as they were >95% similar. The SI sample depth is indicated if not 10 m.

Phylogenetic Relationships Among Gokushovirus MCP Sequences from Saanich Inlet

The phylogenetic relationships among MCP sequences from SI were assessed in relation to MCP sequences from isolates and of other environmental PCR amplicons. The sequenced isolates included phages infecting the parasitic bacteria *Chlamydia* sp., *Bdellovibrio bacteriovorans,* and *Spiroplasma melliferum*; of these only MH2K that infects *B. bacteriovorans* has close marine relatives (**Figure 5**). Marine *Bdellovibrio*-and-like-organisms (BALOs) are commonly found in marine environments and parasitize *Vibrio* sp. (Martin, 2002); hence, viruses similar to MH2K could be infecting marine BALOs. In addition to isolates, phage sequences infecting uncultivated SUP05 from SI obtained using single-cell genomics were also included (Roux et al., 2014), as were MCP sequences from other studies that used metagenomic (Roux et al., 2012b; Yoshida et al., 2013) or targeted amplification approaches (Hopkins et al., 2014).

The majority of the sequences fell within supported (≥90% bootstrap support; ≥4 sequences) phylogenetic groups with four or more sequences (**Figure 5**), although sequences also fell outside these clades. Gokushovirus genomes assembled from metagenomics or single-cell genomics were representative of the SI-5, ENV-2, ENV-4, ENV-9, and WSB-2 clades. Many groups comprised sequences that were location specific (red boxes for the SI groups, dark gray for other environments), indicating that gokushoviruses infect endemic bacteria, and congruent with previous studies that suggest a biogeographic separation of gokushoviruses coupled to specific hosts (Labonté et al., 2009; Tucker et al., 2011). However, many groups contained sequences from multiple environments (light gray boxes on **Figure 5**), suggesting that some gokushoviruses infect widely distributed hosts. In contrast to the 43 unique sequences that were recovered from 77 RFLP patterns observed among 400 analyzed clones (Labonté and Suttle, 2013a), high-throughput sequencing allowed the discovery of five previously unknown groups of gokushoviruses.

Phylogenetic analysis of the 3′ end of the gene encoding the MCP (**Figure 5**) revealed that group SI-5 contained most of the sequences from 200 m, as well as some from 10 m that shared >90% nucleotide identity, but none from other locations. Although all the sequences could not be aligned with the 5′ end of the sequences, members of the SI-5 clade shared >80% pairwise identity with the 5′ end supporting the close phylogenetic relationship among these sequences. The five most common sequences from 200 m, and one of the five most common ones from 10 m, fell within the SI-5 clade (**Figure 5**). Based on the RFLP pattern, sequence SI-13 from an April 2007 anoxic-zone sample from 120 m belonged to the SI-5 group, and was also the fourth most abundant sequence in the 10 m deep-sequencing data. Some of the temporal changes in the gokushovirus phylotypes, such as the presence of RFLP sequence SI-13 in the anoxic zone and 10 m sample in April 2007 likely resulted from changes in the bacterial community. These results agree with a metagenomic study of four aquatic ecosystems, in which the dominant viral taxa persisted over time, while the relative abundances of rare ones constantly changed (Rodriguez-Brito et al., 2010). It has been hypothesized that microbial and viral taxa continuously replace each other in a 'kill-the-winner' manner, maintaining stable metabolic potential and species composition (Rodriguez-Brito et al., 2010). The bacterial community of SI is dynamic and varies based on changing levels of oxygen-deficiency in the water column throughout the year (Zaikova et al., 2010). As SUP05 dominates in the anoxic zone of SI (Zaikova et al., 2010), Group SI-5 sequences are likely from viruses that infect SUP05. In contrast, viruses infecting more ephemeral taxa likely belong to phylogenetic groups with fewer representative OTUs.

The short genetic distance among sequences in Group SI-5 compared to other clades is consistent with a recent infection event. The presence of sequences from both 10 and 200 m suggests that the event occurred in the fall, during deep-water renewal. Interestingly, gokushovirus sequences associated with single-cell genomes from uncultured SUP05 bacteria, found in marine oxygen minimum zones (Roux et al., 2014), fell within the SI-5 gokushoviruses. SUP05 is the most abundant group of bacteria in the anoxic layer of SI (Zaikova et al., 2010), and it seems likely that SI-5 gokushoviruses infect, and are important agents of mortality for SUP05 bacteria and their relatives.

In contrast, the most common sequences from the 10-m sample fell within unsupported groups containing less than five genetically distinct OTUs. An abundant sequence suggests the occurrence of a recent lytic event or an abundant host. Since many sequences did not fall into any phylogenetic group (**Figure 5**), gokushoviruses likely infect a wide range of bacterial taxa, as suggested with other virus groups, where a wider genetic diversity implies a wider diversity of hosts organisms that are infected (Filée et al., 2005; Clasen and Suttle, 2009). Also, the lower richness of gokushoviruses at 200 m than at 10 m in SI (**Figure 4**; **Table 2**) parallels the differences in bacterial richness between depths (Zaikova et al., 2010), suggesting more potential hosts in the oxic than in the anoxic zone. However, the high similarity of the sequences that clustered with the SUP05 phages, and the specific geographic distribution of gokushovirus sequences, suggest that each gokushovirus has a narrow host range.

In general, a power-law distribution of viral taxa indicates an environment in which viruses infect the competitively dominant hosts (Hoffmann et al., 2007). The viral taxa at 200 m were not as well described by a power-law function as those at 10 m (**Figure 2**), likely because there were fewer sequences and the sequences were more similar to each other. Because the five most abundant sequences at 200 m were in group SI-5 and the frequency distribution of the dominant taxa did not fit a power-law function, it is likely that most sequences from 200 m were the result of a recent lytic event of cells within the SUP05 group, as suggested in Roux et al. (2014). Phenomena that could explain the presence of similar sequences at 10 and 200 m are the yearly deep-water renewal and the sinking and resuspension of viral particles. Deep-water renewal does not result in complete mixing of the water column. Rather, the oxygenated water flowing into SI is denser than the basin water; therefore, it sinks and displaces deep-basin waters upward (Anderson, 1973). An

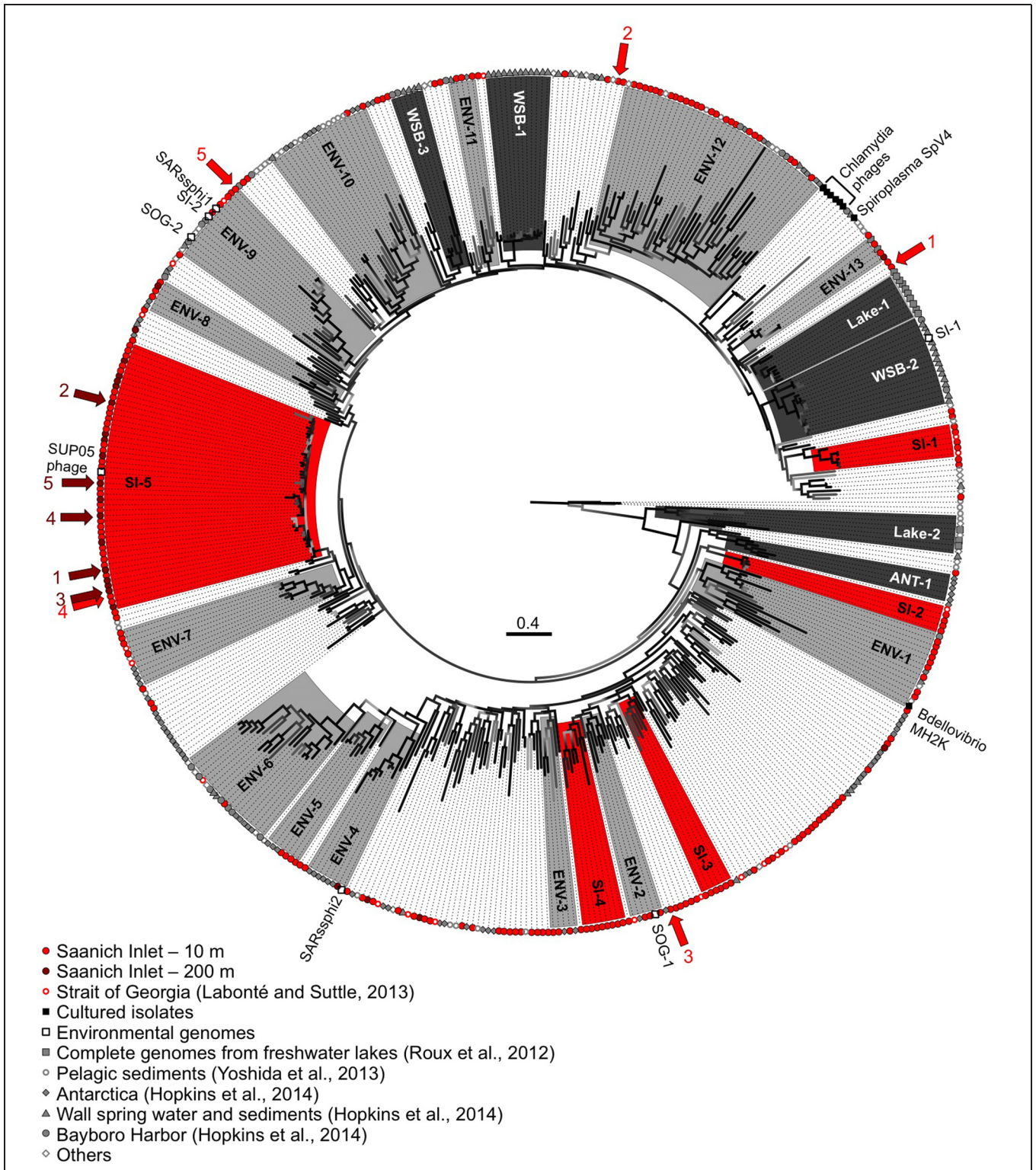


FIGURE 5 | Genetic relatedness of the gokushovirus major capsid protein (MCP) from SI. Unrooted phylogenetic analysis (maximum likelihood; HKY85 model; aLRT probabilities) of PCR products from SI (10 m in red, 200 m in dark red). Bootstrap support is represented by a gradient from light gray to black with black representing supported groups with ≥90% support. Red, dark gray, and light gray shaded phylogenetic groups represent well-supported clades of gokushoviruses with more than five sequences from SI only, a single environment, or multiple environments, respectively. The five most abundant sequences from 10 and 200 m are indicated by the red and dark red arrows, respectively, with the ranking indicated by the associated numbers. The scale bar represents 0.4 nucleotide changes per site.

alternative hypothesis is that viruses which sediment on particles during stratification are transported upward during renewal.

No marine gokushovirus has been isolated so far, but they likely infect a wide range of hosts throughout the water column. A 16S ribosomal RNA gene survey from SI (Zaikova et al., 2010) revealed that Bacteroidetes, δ-proteobacteria (Nitrospina), Actinobacteria (Microthrix), and Verrucomicrobia were more abundant at 10 m than at 200 m, and are potential host taxa. Prophages with a similar genome organization to gokushoviruses have been found in the genomes of bacteria from the phylum Bacteroidetes (Krupovic and Forterre, 2011), supporting the idea that these bacteria may be hosts for gokushoviruses in SI.

This study demonstrated that the genetic richness of gokushoviruses was much higher in the oxic (10 m) than anoxic (200 m) layers of SI, and that a power-law function better described the taxonomic distribution of gokushoviruses at 10 m than 200 m, reflecting the bacterial diversity through the redoxcline. Finally, the presence of very similar viruses at 10 m and 200 m is likely due to deep water renewal or potentially biomass sinking from the surface. These results suggest that gokushoviruses infect a wide range of hosts, but that the host range of an individual genotype is narrow.

Acknowledgments

We thank members of the Suttle laboratory for collecting and processing the samples, and the Hallam laboratory for providing filtered water from Saanich Inlet that made this study possible. We also thank the reviewers for their comments which have resulted in an improved manuscript. This research was supported by the Natural Science and Engineering Research Council of Canada (NSERC) through a postgraduate scholarship (JL) and grants awarded to CS, and SH including NSERC Discovery, Canada Foundation for Innovation (CFI), and the Canadian Institute for Advanced Research (CIFAR). Sample collection was facilitated through ship-time grants from NSERC that supported sample collections in Saanich Inlet (P. D. Tortell and SH). Access to sequencing was funded by the Gordon and Betty Moore Foundation through GBMF1799 to the Broad Institute, and by NSERC and the Tula Foundation using facilities at McGill University and Genome Quebec Innovation Centre.

References

Anderson, J. (1973). Deep water renewal in Saanich Inlet, an intermittently anoxic basin. *Estuar. Coast. Mar. Sci.* 16, 1–10. doi: 10.1016/0302-3524(73)90052-2

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Binga, E. K., Lasken, R. S., and Neufeld, J. D. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J.* 2, 233–241. doi: 10.1038/ismej.2008.10

Breitbart, M., Miyake, J. H., and Rohwer, F. (2004). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS Microbiol. Lett.* 236, 249–256. doi: 10.1111/j.1574-6968.2004.tb09654.x

Brentlinger, K. L., Hafenstein, S., Novak, C. R., Fane, B. A., Borgon, R., McKenna, R., et al. (2002). Microviridae, a family divided: isolation, characterization, and genome sequence of phiMH2K, a bacteriophage of the obligate intracellular parasitic bacterium *Bdellovibrio bacteriovorus*. *J. Bacteriol.* 184, 1089–1094. doi: 10.1128/jb.184.4.1089-1094.2002

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43, 783–791. doi: 10.2307/2531532

Chen, F., Wang, K., Huang, S., Cai, H., Zhao, M., Jiao, N., et al. (2009). Diverse and dynamic populations of cyanobacterial podoviruses in the Chesapeake Bay unveiled through DNA polymerase gene sequences. *Environ. Microbiol.* 11, 2884–2892. doi: 10.1111/j.1462-2920.2009.02033.x

Chipman, P. R., Agbandje-McKenna, M., Renaudin, J., Baker, T. S., and McKenna, R. (1998). Structural analysis of the spiroplasma virus, SpV4: implications for evolutionay variation to obtain host diversity among the *Microviridae*. *Structure* 6, 135–145. doi: 10.1016/S0969-2126(98)00016-1

Clasen, J. L., and Suttle, C. A. (2009). Identification of freshwater phycodnaviridae and their potential phytoplankton hosts, using DNA pol sequence fragments and a genetic-distance analysis. *Appl. Environ. Microbiol.* 75, 991–997. doi: 10.1128/AEM.02024-08

Dean, F., Nelson, J., Giesler, T., and Lasken, R. (2001). Rapid amplification of plasmid and phage DNA using phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* 11, 1095–1099. doi: 10.1101/gr.180501

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., et al. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452, 340–343. doi: 10.1038/nature06735

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

Everson, J. S., Garner, S. A., Fane, B. A., Liu, B. L., Lambden, P. R., and Clarke, I. N. (2002). Biological properties and cell tropism of Chp2, a bacteriophage of the obligate intracellular bacterium *Chlamydophila abortus*. *J. Bacteriol.* 184, 2748–2754. doi: 10.1128/JB.184.10.2748-2754.2002

Filée, J., Tétart, F., Suttle, C. A., and Krisch, H. M. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12471–12476. doi: 10.1073/pnas.0503404102

Garner, S. A., Everson, J. S., Lambden, P. R., Fane, B. A., and Clarke, I. N. (2004). Isolation, molecular characterisation and genome sequence of a bacteriophage (Chp3) from *Chlamydophila pecorum*. *Virus Genes* 28, 207–214. doi: 10.1023/B:VIRU.0000016860.53035.f3

Gobet, A., Böer, S. I., Huse, S. M., van Beusekom, J. E. E., Quince, C., Sogin, M. L., et al. (2012). Diversity and dynamics of rare and of resident bacterial populations in coastal sands. *ISME J.* 6, 542–553. doi: 10.1038/ismej.2011.132

Godson, G. N., Barrell, B. G., Staden, R., and Fiddes, J. C. (1978). Nucleotide sequence of bacteriophage G4 DNA. *Nature* 276, 236–247. doi: 10.1038/276236a0

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Hill, M. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology* 54, 427–432. doi: 10.2307/1934352

Hoffmann, K. H., Rodriguez-Brito, B., Breitbart, M., Bangor, D., Angly, F. E., Felts, B., et al. (2007). Power law rank-abundance models for marine phage communities. *FEMS Microbiol. Lett.* 273, 224–228. doi: 10.1111/j.1574-6968.2007.00790.x

Hopkins, M., Kailasan, S., Cohen, A., Roux, S., Tucker, K. P., Shevenell, A., et al. (2014). Diversity of environmental single-stranded DNA phages revealed by PCR amplification of the partial major capsid protein. *ISME J.* 8, 2093–2103. doi: 10.1038/ismej.2014.43

Huang, S., Wilhelm, S. W., Jiao, N., and Chen, F. (2010). Ubiquitous cyanobacterial podoviruses in the global oceans unveiled through viral DNA polymerase gene sequences. *ISME J.* 4, 1243–1251. doi: 10.1038/ismej.2010.56

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898. doi: 10.1111/j.1462-2920.2010.02193.x

Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

King, A., Adams, M., Carstens, E., and Lefkowitz, E. (2012). "Virus Taxonomy," in *Ninth Report of the International Committee on Taxonomy of Viruses*, 2nd Edn, eds A. King, M. Adams, E. Carstens, and E. Lefkowitz (San Diego, CA: Elsevier Academic Press).

Krupovic, M., and Forterre, P. (2011). Microviridae goes temperate: microvirus-related proviruses reside in the genomes of bacteroidetes. *PLoS ONE* 6:e19893. doi: 10.1371/journal.pone.0019893

Labonté, J. M., Reid, K. E., and Suttle, C. A. (2009). Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. *Appl. Environ. Microbiol.* 75, 3634–3640. doi: 10.1128/AEM.02317-08

Labonté, J. M., and Suttle, C. A. (2013a). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Front. Microbiol.* 4:404. doi: 10.3389/fmicb.2013.00404

Labonté, J. M., and Suttle, C. A. (2013b). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

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

Lilley, M., Baross, J., and Gordon, L. (1982). Dissolved hydrogen and methane in Saanich Inlet, British Columbia. *Deep Sea Res. Part A.* 29, 1471–1484. doi: 10.1016/0198-0149(82)90037-1

Liu, B. L., Everson, J. S., Fane, B. A., Giannikopoulou, P., Vretou, E., Lambden, P. R., et al. (2000). Molecular characterization of a bacteriophage (Chp2) from *Chlamydia psittaci*. *J. Virol.* 74, 3464–3469. doi: 10.1128/JVI.74.8.3464-3469.2000

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A., and Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* 326, 858–861. doi: 10.1126/science.1179287

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380. doi: 10.1038/nature03959

Martin, M. O. (2002). Predatory prokaryotes: an emerging research opportunity. *J. Mol. Microbiol. Biotechnol.* 4, 467–477.

McDaniel, L. D., Rosario, K., Breitbart, M., and Paul, J. H. (2014). Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environ. Microbiol.* 16, 570–585. doi: 10.1111/1462-2920.12184

Polson, S. W., Wilhelm, S. W., and Wommack, K. E. (2011). Unraveling the viral tapestry (from inside the capsid out). *ISME J.* 5, 165–168. doi: 10.1038/ismej.2010.81

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinform.* 12:38. doi: 10.1186/1471-2105-12-38

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F. E., Breitbart, M., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *Water* 4, 739–751. doi: 10.1038/ismej.2010.1

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012a). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* 7:e33641. doi: 10.1371/journal.pone.0033641

Roux, S., Krupovic, M., Poulet, A., Debroas, D., and Enault, F. (2012b). Evolution and diversity of the *Microviridae viral* family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE* 7:e40418. doi: 10.1371/journal.pone.0040418

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. doi: 10.7554/eLife.03125

Salim, O., Skilton, R. J., Lambden, P. R., Fane, B. A., and Clarke, I. N. (2008). Behind the chlamydial cloak: the replication cycle of chlamydiaphage Chp2, revealed. *Virology* 377, 440–445. doi: 10.1016/j.virol.2008.05.001

Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., et al. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio* 5:e01371–e01314. doi: 10.1128/mBio.01371-14

Short, C. M., and Suttle, C. A. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl. Environ. Microbiol.* 71, 480–486. doi: 10.1128/AEM.71.1.480-486.2005

Short, S. M., and Suttle, C. A. (2002). Sequence analysis of marine virus communities reveals that groups of related algal viruses are widely distributed in nature. *Appl. Environ. Microbiol.* 68, 1290–1296. doi: 10.1128/AEM.68.3.1290-1296.2002

Smith, R. J., Jeffries, T. C., Roudnew, B., Seymour, J. R., Fitch, A. J., Simons, K. L., et al. (2013). Confined aquifers as viral reservoirs. *Environ. Microbiol. Rep.* 5, 725–730. doi: 10.1111/1758-2229.12072

Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., et al. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci. U.S.A.* 103, 12115–12120. doi: 10.1073/pnas.0605127103

Storey, C. C., Lusher, M., and Richmond, S. J. (1989). Analysis of the complete nucleotide sequence of Chp1, a phage which infects avian *Chlamydia psittaci*. *J. Gen. Virol.* 70, 3381–3390. doi: 10.1099/0022-1317-70-12-3381

Sullivan, M. B., Coleman, M. L., Quinlivan, V., Rosenkrantz, J. E., Defrancesco, A. S., Tan, G., et al. (2008). Portal protein diversity and phage ecology. *Environ. Microbiol.* 10, 2810–2823. doi: 10.1111/j.1462-2920.2008.01702.x

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

Suttle, C. A., Chan, A. M., and Cottrell, M. T. (1991). Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl. Environ. Microbiol.* 57, 721–726.

Tucker, K. P., Parsons, R., Symonds, E. M., and Breitbart, M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J.* 5, 822–830. doi: 10.1038/ismej.2010.188

Ward, B., Kilpatrick, K., Wopat, A., Minnich, E., and Lidstrom, M. (1989). Methane oxidation in Saanich Inlet during summer stratification. *Cont. Shelf. Res.* 9, 65–75. doi: 10.1016/0278-4343(89)90083-6

Yoshida, M., Takaki, Y., Eitoku, M., Nunoura, T., and Takai, K. (2013). Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS ONE* 8:e57271. doi: 10.1371/journal.pone.0057271

Zaikova, E., Walsh, D. A., Stilwell, C. P., Mohn, W. W., Tortell, P. D., and Hallam, S. J. (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ. Microbiol.* 12, 172–191. doi: 10.1111/j.1462-2920.2009.02058.x

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 16 March 2015
doi: 10.3389/fmicb.2015.00199

Genomic characteristics and environmental distributions of the uncultivated Far-T4 phages

Simon Roux[1]*, *François Enault*[2,3], *Viviane Ravet*[2,3], *Olivier Pereira*[2,3] and *Matthew B. Sullivan*[1]*

[1] Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, USA, [2] Laboratoire "Microorganismes: Génome et Environnement," Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France, [3] Centre National de la Recherche Scientifique, UMR 6023, Laboratoire Microorganismes: Génome et Environnement, Aubière, France

Viral metagenomics (viromics) is a tremendous tool to reveal viral taxonomic and functional diversity across ecosystems ranging from the human gut to the world's oceans. As with microbes however, there appear vast swaths of "dark matter" yet to be documented for viruses, even among relatively well-studied viral types. Here, we use viromics to explore the "Far-T4 phages" sequence space, a neighbor clade from the well-studied T4-like phages that was first detected through PCR study in seawater and subsequently identified in freshwater lakes through 454-sequenced viromes. To advance the description of these viruses beyond this single marker gene, we explore Far-T4 genome fragments assembled from two deeply-sequenced freshwater viromes. Single gene phylogenetic trees confirm that the Far-T4 phages are divergent from the T4-like phages, genome fragments reveal largely collinear genome organizations, and both data led to the delineation of five Far-T4 clades. Three-dimensional models of major capsid proteins are consistent with a T4-like structure, and highlight a highly conserved core flanked by variable insertions. Finally, we contextualize these now better characterized Far-T4 phages by re-analyzing 196 previously published viromes. These suggest that Far-T4 are common in freshwater and seawater as only four of 82 aquatic viromes lacked Far-T4-like sequences. Variability in representation across the five newly identified clades suggests clade-specific niche differentiation may be occurring across the different biomes, though the underlying mechanism remains unidentified. While complete genome assembly from complex communities and the lack of host linkage information still bottleneck virus discovery through viromes, these findings exemplify the power of metagenomics approaches to assess the diversity, evolutionary history, and genomic characteristics of novel uncultivated phages.

Keywords: T4 phages, freshwater ecology, Caudovirales, capsid proteins, viral genomes

OPEN ACCESS

Edited by:
Alejandro Reyes, Universidad de los Andes, Colombia

Reviewed by:
André M. Comeau, Dalhousie University, Canada
Jonathan Filée, Centre National de la Recherche Scientifique, France

***Correspondence:**
Simon Roux and Matthew B. Sullivan, Department of Ecology and Evolutionary Biology, University of Arizona, 1007 E. Lowell St., Tucson, AZ 85719, USA
simroux@email.arizona.edu; mbsulli@email.arizona.edu

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *01 December 2014*
Accepted: *24 February 2015*
Published: *16 March 2015*

Citation:
Roux S, Enault F, Ravet V, Pereira O and Sullivan MB (2015) Genomic characteristics and environmental distributions of the uncultivated Far-T4 phages. Front. Microbiol. 6:199. doi: 10.3389/fmicb.2015.00199

Introduction

Viruses are the most abundant biological entities in the biosphere, impact microbial communities structure, alter cellular genomes evolutionary history and indirectly influence major biogeochemical cycles (Breitbart et al., 2007; Suttle, 2007; Rohwer et al., 2009; Hurwitz et al., 2013). Despite these important roles, most viruses in nature (~75–95%) remain uncharacterized in any well-sampled

viral community (Brum et al., in press). While new isolates will certainly help with this –e.g., viruses for two of the most abundant marine bacteria (SAR11 and SAR1116) and rare virosphere representatives (infecting *Cellulophaga*) were described only last year (Holmfeldt et al., 2013; Kang et al., 2013; Zhao et al., 2013), rapid means of discovering and characterizing viruses are needed. One approach is to pull viral signals out of microbial genomic datasets, with single-cell amplified genomes (SAGs) likely offering the best hope of obtaining complete viral genomes for uncultivated hosts (e.g., Roux et al., 2014a). Another is to use assembled genomic contigs from viral metagenomes (viromes) to explore the genomic context for novel viral groups (e.g., Emerson et al., 2012; Minot et al., 2012a,b; Dutilh et al., 2014).

Among viruses infecting bacteria (bacteriophages or phages), the T4-like superfamily (*Tevenvirinae*) is one of the most widespread, abundant, and extensively studied group. The *Tevenvirinae* are members of the *Myoviridae* order, tailed bacteriophages with a double-stranded DNA genome, and were first isolated and characterized on *Escherichia coli* (Miller et al., 2003b). Other members of this superfamily were subsequently isolated on *Aeromonas* (Petrov et al., 2010; Kim et al., 2012), *Vibrio* (Miller et al., 2003a), *Prochlorococcus* and *Synechococcus* (Sullivan et al., 2010), and *Pelagibacter* (Zhao et al., 2013).

The abundance of T4 phages in natural communities, largely assessed by marker genes, has been the subject of significant effort since initial PCR-based analyses were implemented in 1998 (Fuller et al., 1998). Subsequent studies, targeting the portal protein (T4 phage gene 20) and major capsid protein (MCP, T4 phage gene 23) genes, ensued across marine (Millard et al., 2004; Filée et al., 2005; Zeidner et al., 2005; Sullivan et al., 2006, 2008; Sharon et al., 2007; Comeau and Krisch, 2008; Goldsmith et al., 2011), and freshwater (Dorigo et al., 2004; Chénard and Suttle, 2008; Butina et al., 2010; Matteson et al., 2011; Hewson et al., 2012) samples. While criticized as a means to quantitatively evaluate T4 phage ecology (Sullivan et al., 2008; Duhaime and Sullivan, 2012; Sullivan, 2015), such marker gene surveys have clearly helped document the diversity of T4 phage marker genes and establish hypotheses about evolutionary history and taxonomy in wild T4 phages. Specifically, the *Tevenvirinae* appear comprised of several subgroups including (i) the "true" T-evens represented by T4 and closely related phages infecting *Enterobacteria* (e.g., T2, T6), (ii) the Pseudo T-evens and Schizo T-evens (including *Aeromonas* and *Vibrio* phages), morphologically distinguishable, and (iii) the more distant Exo T-evens (including cyano- and pelagiphages).

Beyond marker genes, the T4 phage group has also been relatively extensively explored at the whole genome level. A "core-genome" shared across all or most members of the *Tevenvirinae* was defined, representing functions like DNA replication, repair and recombination, virion morphogenesis or control of gene expression (Sullivan et al., 2005, 2010; Petrov et al., 2010). Further, hierarchical "core" gene sets from subsets of these phages and flexible genes sporadically distributed across these genomes suggested means by which T4 phages differentiate to different environments and hosts (Millard et al., 2004; Mann et al., 2005; Weigele et al., 2007; Petrov et al., 2010; Sullivan et al., 2010). The largely similar genome organization and predominantly vertical evolutionary history of core genes hint at robust taxonomic boundaries in this phage group (Ignacio-Espinoza and Sullivan, 2012), and recent exploration of genomic variability in wild T4-like cyanophages confirmed such discrete structure in sequence space and empirically placed limits between populations at about 95% nucleotide identity (Deng et al., 2014).

T4-like phage sequences were also mined from the Global Ocean Sampling (GOS) expedition microbial metagenomic dataset (i.e., the viral signal here originate from actively infected cells captured on filters) to design new degenerate PCR primers which revealed a new T4 phage group—the "Far-T4" phages (Comeau and Krisch, 2008). This clade includes a very peculiar phage: RM378, isolated on the thermophilic bacterium *Rhodothermus marinus* (Hjorleifsdottir et al., 2014). Morphologically, phage RM378 is similar to a T4-like phage (A2 morphology) and encodes a T4-like capsid protein gene, but its genome contains only half of the 38 (core) genes conserved in 26 T4-like phage genomes available for comparative study (Sullivan et al., 2010). Moreover, the RM378 genome lacks a readily identifiable structural or replication module that is discernible among all other *Tevenvirinae*. Far-T4 phage major capsid proteins have since then been detected in marine (Williamson et al., 2012; Hurwitz and Sullivan, 2013) and freshwater (Roux et al., 2012) viromes, but no formal genomic evaluation of the Far-T4 phages is available beyond the reference RM378 genome (Hjorleifsdottir et al., 2014).

Here, to expand our understanding of Far-T4 phages, we assembled genome fragments from two deeply-sequenced freshwater viromes, and used these to evaluate the evolutionary history of the Far-T4 phages and of their major capsid protein, as well as assess their global distribution in freshwater and marine ecosystems using 196 previously published viromes.

Results

Detection of Far-T4 Contigs

Reads from 2 deeply-sequenced viromes from the Lake Pavin (sampled at 4 and 8 m) and 2 previously published 454 viromes from surface samples of Lakes Pavin and Bourget (Roux et al., 2012) were assembled into genome fragments and searched for *g23* genes. Overall, 32 Far-T4 *g23* genes were detected in the two deeply-sequenced viromes, and eight in the 454 viromes (**Table 1**). Using these and publicly available sequences, the diversity and structure of the Far-T4 phages was evaluated using a Gp23-based phylogenetic tree (**Figure 1**). This tree clearly resolves the T4-like phages ("Near-T4") from the T4-like cyanophages ("Cyano-T4"), along with two recently described *Alphaproteobacteria* phages (infecting SAR11 and *Sinorhizobium*) and the Far-T4 phages.

These latter sequences form a monophyletic group composed of (i) *Rhodothermus* phage RM378, the only reference Far-T4 genome available (in black), (ii) the PCR-amplified sequences from seawater used to first define the Far-T4 group (in red), (iii) 8 sequences retrieved from 454-sequenced and published freshwater lakes viromes (in green), and (iv) 32 sequences from the 2 viromes analyzed here sampled from Lake Pavin (in light and dark blue). Within this monophyletic Far-T4 group, five

TABLE 1 | List of Far-T4 contigs assembled from freshwater viromes.

Dataset	Sequence Id	Length	Clade	Marker genes	PhoH	Putative host group (CRISPR)	Putative host family (tetranucleotide)
HiSeq Viromes	**Pavin_2013_4m_8**	**105162**	**Far_T4_1**	***g23; g20; g17***	+	*Clostridia ?*	*Anaplasmataceae ?*
	Pavin_2013_4m_10	**97022**	**Far_T4_3**	***g23; g20; g17***	+		
	Pavin_2013_4m_12	86378	Far_T4_1	*g23; g20; g17*	+		
	Pavin_2013_8m_12	59230	Far_T4_1	*g23; g20; g17*	+		
	Pavin_2013_4m_33	55178	Far_T4_1	*g23; g20; g17*	+		
	Pavin_2013_4m_40	51859	Far_T4_1	*g23; g20; g17*	+		*Anaplasmataceae ?*
	Pavin_2013_8m_18	51839	Far_T4_1	*g23; g20; g17*	+		*Anaplasmataceae ?*
	Pavin_2013_4m_55	43868	Far_T4_1	*g23; g20; g17 (partial)*	+		
	Pavin_2013_4m_62	**40954**	**Far_T4_5**	***g23***			
	Pavin_2013_4m_77	**36429**	**Far_T4_4**	***g23***	+		
	Pavin_2013_8m_34	35829	Far_T4_4	*g23*	+		
	Pavin_2013_8m_73	27012	Far_T4_4	*g23*	+		
	Pavin_2013_4m_174	23723	Far_T4_4	*g23*	+		
	Pavin_2013_4m_229	20766	Far_T4_1	*g23*	+		
	Pavin_2013_4m_232	20663	Far_T4_4	*g23; g20; g17*			
	Pavin_2013_8m_128	19577	Far_T4_1	*g23*	+		
	Pavin_2013_8m_153	17698	Far_T4_4	*g23*	+		
	Pavin_2013_4m_295	17576	Far_T4_3	*g23*			
	Pavin_2013_8m_177	16514	Far_T4_4	*g23*	+		
	Pavin_2013_4m_335	16154	Far_T4_1	*g23; g20; g17*	+	*Bacilli ?*	
	Pavin_2013_4m_405	14371	Far_T4_1	*g23*	+		
	Pavin_2013_8m_232	14013	Far_T4_1	*g23; g20; g17 (partial)*	+		
	Pavin_2013_4m_436	13727	Far_T4_1	*g23*	+		
	Pavin_2013_4m_473	13214	Far_T4_5	*g23*			
	Pavin_2013_8m_261	13214	Far_T4_5	*g23*			
	Pavin_2013_4m_512	12683	Far_T4_1	*g23*	+		
	Pavin_2013_4m_546	12156	Far_T4_1	*g23*	+		
	Pavin_2013_8m_320	11847	Far_T4_5	*g23*			
	Pavin_2013_8m_328	11687	Far_T4_1	*g23; g20; g17*		*Bacilli ?*	
	Pavin_2013_4m_634	10875	Far_T4_4	*g23*			
	Pavin_2013_4m_647	10681	Far_T4_4	*g23; g20*			
	Pavin_2013_4m_691	10076	Far_T4_1	*g23; g20*			
454 Viromes	Pavin_2009_08187	5217	Far_T4_1	*g23*			
	Bourget_2009_19709	2282	Far_T4_1	*g23*	+		
	Bourget_2009_14420	3756	Far_T4_1	*g23*	+		
	Bourget_2009_12531	2307	Far_T4_1	*g23*	+		
	Bourget_2009_12442	2049	Far_T4_1	*g23*	+		
	Pavin_2009_09564	1527	Far_T4_4	*g23*			
	Bourget_2009_57	1630	Far_T4_4	*g23*			
	Bourget_2009_02710	2702	Far_T4_1	*g23*			

Contigs selected as reference for each clade are highlighted in bold.

clades can be robustly delineated (bootstraps >80%) including (i) a clade of both seawater amplicons and freshwater virome-derived sequences that represents the majority of the sequences (Far-T4 clade 1), (ii) a clade of seawater amplicons only (Far-T4 clade 2), and (iii) three clades composed solely of freshwater virome sequences (Far-T4 clades 3, 4, and 5). Notably, the Gp23 sequence of Rhodothermus marinus phage RM378 is clearly distinct from all other Far-T4 phage group members, so its usefulness as a reference genome for the Far-T4 group may be limited.

Phylogenetic trees computed from two other phylogenetic markers—the portal protein (Gp20) and large terminase subunit (Gp17) – confirmed the robust and well-supported monophyly of the Far-T4 phages and the delineation of clades 1, 3, 4, and 5 (Figure S1, only *g23* amplicons are available for clade 2, which could thus not be detected with another marker gene).

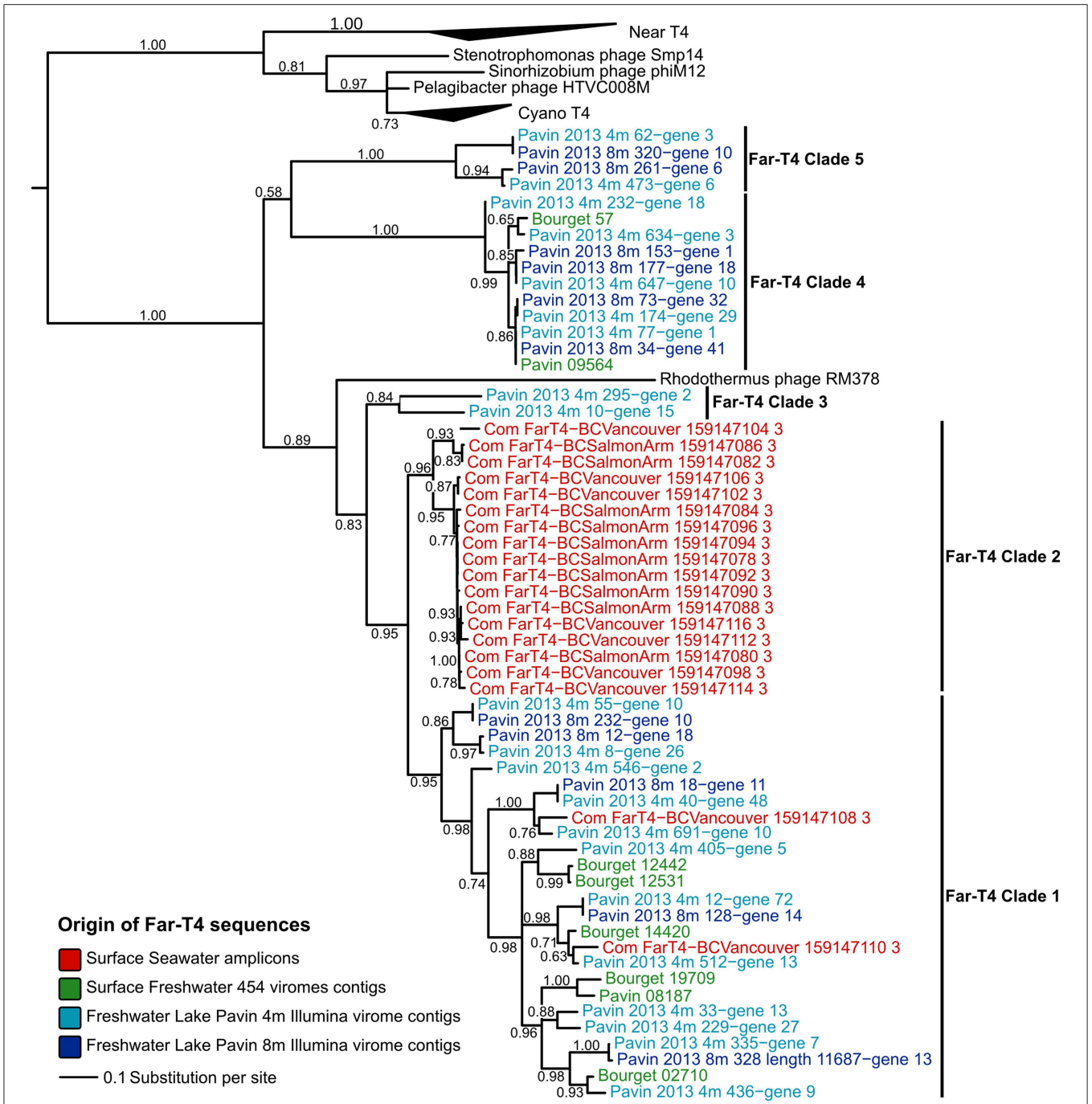


FIGURE 1 | Phylogenetic tree computed from Gp23 (major capsid protein) multiple alignment. As this tree includes the PCR sequences used to define the Far-T4 group, the multiple alignment was trimmed to include only positions matching these amplicons. All nodes with bootstraps lower than 50 were collapsed. Far-T4 sequences are color-coded according to their original sample, and reference sequences are in black.

Insights into Far-T4 Gene Content and Genome Organization

We next used the 12 Far-T4 contigs longer than 25 kb to evaluate the genome content and organization of these new phages. As with all *Tevenvirinae* except RM378, clear preservation of gene order and functional modularity could be delineated (**Figure 2**). These include structural genes (e.g., portal, terminase, and tail genes) proximal to the major capsid protein (MCP), and replication genes located in a module elsewhere in the genome including DNA polymerases, primases, helicases and exonucleases (**Figure 2**). Surprisingly, nearby to the structural genes was also a *phoH* gene, which in *E coli* represents a phosphate

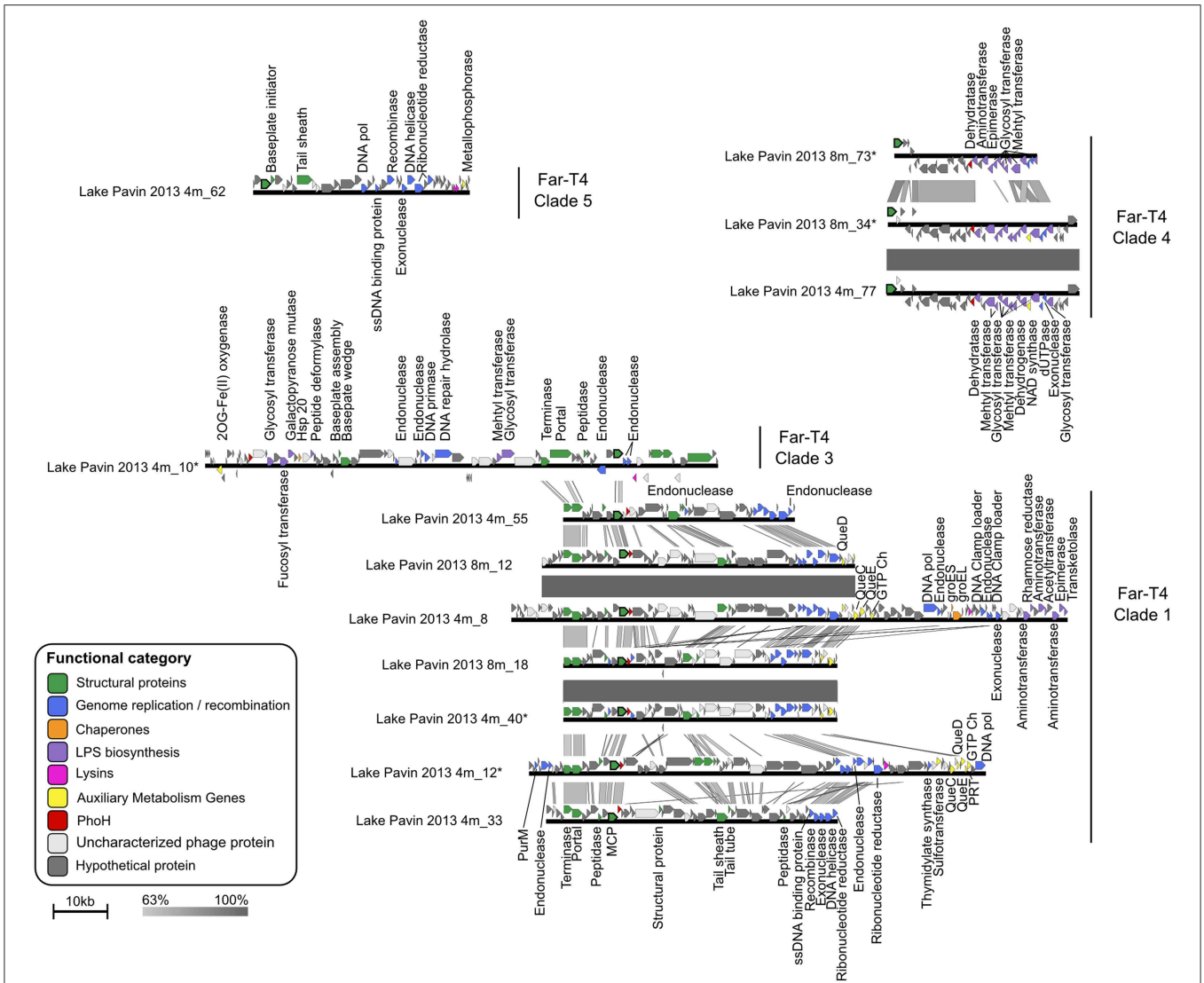


FIGURE 2 | Comparison of Far-T4 contigs assembled from the Lake Pavin viromes. The two deeply sequenced viromes from Lake Pavin allowed for the assembly of large genome fragments, which are here compared in terms of gene similarity and order. Genes are color-coded according to their functional annotation, and major capsid genes (g23) are outlined in black. Significant sequence similarity between successive sequences are highlighted by gray squares, shaded according to percentage of amino-acid identity. Predicted proteins with no associated function are classified as "Uncharacterized phage proteins" when at least one similar sequence was detected in a known phage and "Hypothetical protein" otherwise. PurM, Phosphoribosylformylglycinami-dine cyclo-ligase; GTP Ch, GTP cyclohydrolase; DNA Pol, DNA polymerase; PRT, Phosphoribosyl transferase; Hsp, Heat shock protein.

starvation-induced ATPase (Kim et al., 1993). The *phoH* gene is "core" to marine T4-like cyanophages (Sullivan et al., 2010), has been documented across a wide variety of phages, and used as a marker gene to assess marine phage diversity (Goldsmith et al., 2011), but its function as a phosphate-stress related gene outside of *E. coli* remains controversial (Sullivan et al., 2010). Taken together, the facts that a *phoH* gene occurs in 3 of 4 Far-T4 phage clades for which genome fragments are available (**Table 1**), is detected near to the MCP for 2 clades (Far-T4 1 and 4), and that trees from PhoH are consistent with the other virion-associated marker genes (Figure S1) suggests strong conservation and likely an important function for this gene in freshwater phages.

Alongside these more typical viral genes, Far-T4 contigs from clade 1 also harbor several Auxiliary Metabolic Genes [AMGs, *sensu* (Breitbart et al., 2007)], notably from the queuosine biosynthesis pathway classifed as class II AMGs (Hurwitz et al., 2015). Queuosine (Que) is a hyper-modified guanosine used in tRNAs specific for four amino acids (Asp, Asn, His, or Tyr) and found across all domains of life (El Yacoubi et al., 2012). This detection of *Que* genes in freshwater Far-T4 complements the recent description of near-complete Que biosynthesis pathway in

phages infecting virulent *Streptococcus pneumoniae* strains (Sabri et al., 2011), as well as marine *Cellulophaga baltica* (Holmfeldt et al., 2013), the latter being also detected in in diverse aquatic viromes (55 of 137 screened, Holmfeldt et al., 2013). Taken together, these detections of *Que* genes in phage genomes sampled from such different ecosystems (seawater, freshwater and human lung) suggest a general role in phage cycle for these genes. Interestingly, Sabri et al. (2011) suggested that *Que* genes could act as a feedback signal to control the quantity of phage structural gene transcripts, an hypothesis that would be consistent with the location of these genes in the structural module in Far-T4 contigs.

We next evaluated the diversity and novelty of genes within all Far-T4 contigs. On average, clade 1 genome fragments are the most closely related to database representatives, ∼80% of genes have a hit against the NCBI NR database, whereas this frequency is only ∼70, 60, and 47% for clades 5, 4, and 3 respectively (Figure S2). Of the novel genes (not hitting anything in databases) genes, 60 to 100% remain conserved within a clade for clades 1, 4, and 5. In contrast, most (80%) of the novel genes in clade three remain unique to each contig and appear not to be conserved within this clade, even though the contigs cover the same genome region including the MCP. This difference in genome content associated with the long distance separating these two sequences on the MCP tree (**Figure 1**) and the lower bootstrap support for this clade compared to other Far-T4 clades (84% vs. 95–100%) suggest that these sequences may actually represent two neighbor clades rather than one single group.

Despite the fact that these contigs represent incomplete genome fragments of Far-T4 phages, a clustering based on the proportion of genes shared between pairs of contig consistently recovered the clades established using phylogenetic markers (Figure S3, Robinson-Foulds distance between topologies = 30, the same distance between 1000 randomly permuted trees has an average of 51.4, p-value $< 10^{-16}$). This indicates that contigs from the same clade indeed share more genes between themselves than with other Far-T4 phages. Accordingly, similarity at the nucleotide level is mostly restricted within each clade and barely detectable between Far-T4 clades (**Figure 2**). Finally, this comparison at the nucleotide level highlighted identical contigs separately assembled in the 4 and 8 m samples for each clades one and four, suggesting that virome assembly produced consistent results.

Conservation and Evolution Patterns of the Major Capsid Protein (MCP)

Given the importance of the major capsid protein (MCP) for evaluating the evolutionary history of viruses (Hendrix, 2002; Bamford et al., 2005; Brüssow, 2009; Abrescia et al., 2012), we next extracted the complete MCP sequences (as opposed to the partial amplicons previously available) from our Far-T4 contigs. The Far-T4 phages MCP was previously noted as "very divergent from the rest of the known sequences" (Comeau and Krisch, 2008), which can be linked to an ancient separation between the Far-T4 and the other T4-like phages, or to a relaxed selection pressure on the MCP in the Far-T4 lineage driven by phage-host coevolution dynamics (Hall et al., 2011).

Broadly speaking, Far-T4 MCP sequences are ∼15% longer than their Near-T4 and Cyano-T4 counterparts for all clades (Figure S4). Evolutionarily, the ratio of non-synonymous to synonymous mutations (dN/dS) across the alignment (all T4) is low (0.104 as estimated by PAML; Yang, 2007), which corresponds to a strong stabilizing selection as expected for a functionally important and conserved gene. When allowing for different dN/dS for the Far-T4 phages and the other *Tevenvirinae*, the ratio was only slightly higher in the Far-T4 phages (0.118 vs. 0.093, Table S1). This suggests a strong conservation of the MCP overall, and a sequence divergence between Far-T4 and other T4 phages MCPs likely linked to an ancient separation rather than differences in phage-host interactions.

However, 20 out of 302 sites appeared to be under relaxed selection (dN/dS = 1, p-value 8.7e-102, Table S1), and corresponded to less-conserved residues between highly conserved regions with predicted secondary structures (Figure S5). To investigate this further, we next built 3D models from our assembled Far-T4 MCP, based on the characterized structure common to the T4 phage MCP and vertex protein. This model suggests the following organization: the N- and C-terminal conserved domains are gathered within a "core" conserved region which includes the predicted secondary structures (blue parts on **Figure 3**), flanked by more variable and unstructured parts (i.e., no predicted alpha helix or beta strands) on the outside. Similar folding was predicted for the different Far-T4 clades (Figure S6). It is thus tempting to speculate that the conserved and structured parts are responsible for the core virion structure, and that the more variable parts outside are linked to virion-specific decorations as for the known T4-like MCP structure (Comeau and Krisch, 2008).

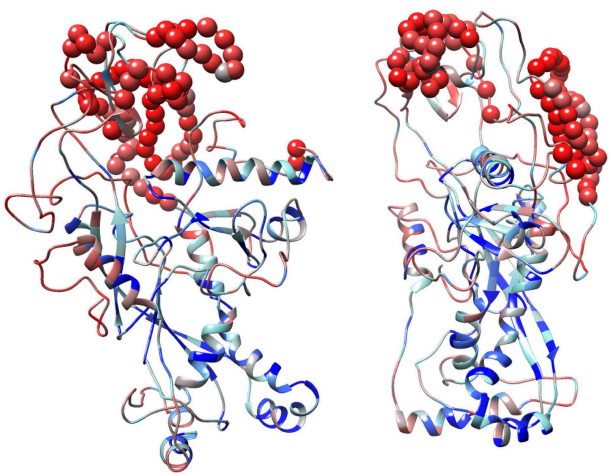


FIGURE 3 | Predicted structure of the Far-T4 major capsid protein (two views rotated by 90° through the y-axis). This model is based on the sequence from contig Pavin_2013_4m_8, and was predicted with I-Tasser based on the structure of the T4 phage vertex protein. Models are colored according to their conservation across T4-like phages, from least (red) to most (blue) conserved. Residues corresponding to insertions in the Far-T4 sequence not present in other T4 phages are highlighted with spheres.

Distribution and Abundance of Far-T4 in Aquatic Environments

To evaluate the distribution and relative abundance of Far-T4 phages in nature, we next used our new reference genome fragments for recruitment analysis against 186 publicly available viromes derived from marine (62), freshwater (26), hypersaline (13) or eukaryote-associated (85) samples. To consider a Far-T4 phage as being "present" in a virome, we required at least 100 reads to be recruited per genome fragment (blastp *e*-value < 0.001, score > 50, see Materials and Methods). Overall, Far-T4 phages were detected in 15 freshwater (from López-bueno et al., 2009; Rosario et al., 2009a; Roux et al., 2012; Ge et al., 2013; Tseng et al., 2013) and 37 seawater viromes (from Williamson et al., 2008, 2012; Hurwitz and Sullivan, 2013), or ∼60% of all temperate aquatic viromes (i.e., not hypersaline), which included samples from the Pacific, Indian and Atlantic oceans as well as lakes in Europe, Asia, Antarctica and North America (Table S2). All but one of these viromes included sequences similar to every Far-T4 clade (clade five was not detected in the Spring sample from Lake Limnopolar), which suggests that the whole Far-T4 group is relatively widespread among aquatic environments.

Refining these analyses to require non-redundant recruitment to the newly available Far-T4 phage contigs suggested that clades one and four were more prevalent in freshwater viromes (*t*-test *p*-values $< 10^{-07}$), and clades three and five not significantly differently covered between freshwater and seawater (**Figure 4**). The nucleotide identity of the recruited reads was on average 70% for all clades, with up to 100% matches from some freshwater viromes (Figure S7). As previously interpreted (e.g., Holmfeldt et al., 2013) and given what is known about wild T4 cyanophage populations (>95% Average Nucleotide Identity within a population, Deng et al., 2014), this suggests that phages related to Far-T4 are likely occurring in seawater, but the contigs assembled from Lake Pavin represent freshwater-specific populations.

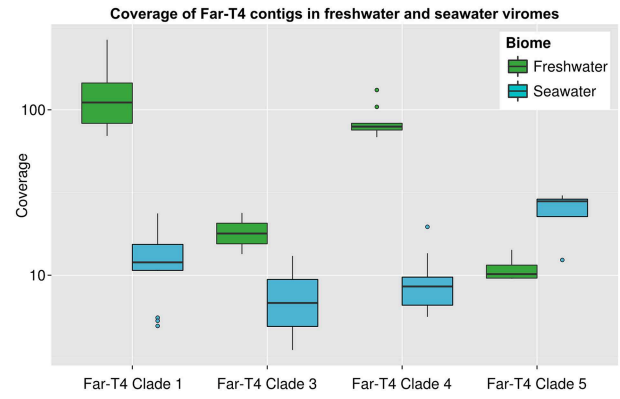


FIGURE 4 | Coverage of different Far-T4 clades in seawater and freshwater viral metagenomes. For each category of virome (freshwater and seawater), the coverage of each contig (i.e., number of bp similar to a contig normalized by the contig length) was computed from 2.5 millions of randomly selected reads from each virome. Coverage could not be assessed for Far-T4 clade 2 since no contigs are available for this clade.

Assessing Putative Host(s) for some Far-T4 Phages

Without a close isolated reference, assessing the putative host(s) of a virus only detected in metagenomes is often difficult. Currently, the most straight-forward *in silico* approach is to look for sequences similar to the newly described virus in microbial genomic datasets, either as prophage (viral genome integrated in the host's genome), as separate contig in a single-amplified genome from an infected cell, or as a CRISPR spacer (i.e., 15–50 bp viral sequence(s) stored in a microbial genome from past infections). Here, no sequence closely related to Far-T4 phages could be detected in public databases of microbial genomes: the most similar genome sequences display only 50–60% amino acid identity (hits from genomic fragments in *Proteobacteria* SAGs, NCBI gis 655454702 and 655453257), and best hits to CRISPR spacers still displayed 2 mismatches when only 100% matches can be trusted on such short nucleotide sequences (hits to a *Clostridia*: *Peptoclostridium difficile*, gi 484228666, and a *Bacilli*: *Streptococcus pneumoniae*, gi 452723578). Thus, no putative host could be predicted for Far-T4 phages based on a search of microbial genomic datasets.

Another approach available is to use genome composition similarities between viral genome(s) and reference microbial genomes (especially tetranucleotide frequencies, Pride et al., 2006). In the Far-T4 case, 3 sequences from the Far-T4 clade 1 display tetranucleotide frequencies close to the small *Alphaproteobacteria Ehrlichia chaffeensis* (**Table 1**), an obligate intracellular parasite mostly found in animals and ticks. Based on a previous evaluation on more than 15,000 known virus-host pairs, such similarity in tetranucleotide frequencies correspond to host family predictions 88 to 98% accurate, and in that case would link Far-T4 phages with small *Alphaproteobacteria* from the *Anaplasmataceae* family.

Discussion

Because so few environmental microbes can be cultivated in the lab (Rappé and Giovannoni, 2003), most viruses infecting these microbes are still to be characterized. In the absence of isolated host, groups such as the Far-T4 phages, although seemingly abundant, could until recently only be studied through single marker genes analyses, either from short-read metagenomes (Roux et al., 2012) or PCR amplification (Comeau and Krisch, 2008). Here, HiSeq Illumina sequencing provided large genome fragments to expand our genomic context and understanding of this Far-T4 group. Specifically, these data help to (i) validate the existence and distribution of several Far-T4 clades based on multiple genes, (ii) evaluate genome organization in this group, and (iii) witness patterns of evolution on conserved genes.

The Far-T4 group, initially identified from marker genes, appears here to represent a set of phages that display T4 phage characteristics, but are distant from the known T4 phages by every marker evaluated. A comparison between Far-T4 contigs further validates the clade delineation based on single marker genes, and suggests that Gp23 is a very good marker for this extended T4 family. Genome comparison also revealed an important genetic diversity within this group: even if all Far-T4 contigs

form a monophyletic clade, they only share a few "core" genes. The consistency in clade delineation using phylogenetic markers and gene content analysis associated with the overall conserved modular genome structure indicates that all Far-T4 likely derived from a (likely distant) common ancestor, and most of the "variable" genes (genes outside of the highly conserved core genes) have either diverged too much to recover any similarity or were subject to genome re-arrangement and/or horizontal gene transfer.

From the genome context perspective, these newly available Far-T4 phage genome fragments clearly evidence that *Rhodothermus marinus* phage RM378, the closest cultured representative of the Far-T4 phages, seems anomalously representative of the Far-T4 at best. Specifically, the Far-T4 phages have similar genomic organizational properties to the Near- and Cyano-T4 phages, while phage RM378 does not. Far-T4 phage genome fragments also display a handful of the T4 "core" genes (10 out of 38), including all major virion-related proteins (notably the major capsid, portal, terminase, and tail proteins) and main replication proteins (DNA polymerase, primase/helicase, and both subunits of ribonucleotide reductase). This suggests that Far-T4 phages will harbor T4-like morphology (as does RM378) and probably T4-like replication mechanism. However, the absence of detection of T4 "core" genes linked to virus-host interactions and transcription regulation indicates that Far-T4 host interaction dynamics, transcription patterns, and infection cycle might differ from the T4 phages, and that their characterization will require further experiments beyond sequence analysis. Alternatively, some T4 "core" genes might also be missing because the Far-T4 contigs represent only partial genomes.

Finally, metagenomic fragment recruitment analyses on these new genomic fragments establish these Far-T4 phages as being widely distributed around the world in aquatic systems—freshwater and seawater. Interestingly, Far-T4 phages were thought to be absent from freshwater systems due to the lack of PCR-based amplicons using newly optimized primer sets (Comeau and Krisch, 2008), however this directly reflects the fact that the PCR primers were designed from seawater sequences, and are not able to amplify Far-T4 sequences assembled from freshwater. This difficulty in designing universal primers, even for specific target groups, is relatively well known in microbial ecology and a source of controversy and debate in trying to establish quantitative viral ecology (Sullivan et al., 2008), and viromics may better represent viral abundances, at least for dsDNA phages (Duhaime and Sullivan, 2012; Sullivan, 2015).

Stepping back, this and related studies that leverage viromics to characterize new viruses (e.g., Rosario et al., 2009a; Emerson et al., 2012; Dutilh et al., 2014) help illustrate the inferential advances and remaining challenges of the approach. Specifically, viromics clearly enables assembly of new phage groups that have eluded cultivation so far, as well as fragment recruitment analytical capabilities to provide environmental context for newly available reference genomes. Yet two main challenges remained: (i) assembling complete and accurate genomes from complex communities, and (ii) extracting information beyond this genome sequence, especially the host(s), quantification in different ecosystems, and characterization of infection cycle of the newly described virus, all required to really assess the potential impact of a virus on ecosystems.

Rigorously evaluating the quality of metagenomic assemblies (i.e., do contigs represent real consensus genomes or *in silico* generated chimeras) remains fundamentally problematic especially since no gold standard metrics (e.g., what is real?) are readily apparent with newly discovered environmental data (Charuvaka and Rangwala, 2011; Luo et al., 2012; Vázquez-Castellanos et al., 2014). For the most abundant viruses, the high coverage provided by most recent sequencing technologies seems to lead to accurate and reproducible assemblies, as testified by PCR confirmation of metagenome-based assemblies (Dutilh et al., 2014) or the recovery of identical contigs from separate samples in this study. However, such high coverage is not yet available for members of the "rare virosphere." Several workarounds have been proposed to access this rare virosphere such as single cell viromics (Allen et al., 2011), viral tagging (Deng et al., 2014) or targeted viromics (Brum et al., 2013). Additionally, even if robust assemblies of large genome fragments are now possible, the assembly of complete genomes from complex communities is still relatively rare, notably because of the presence of repeat regions and highly conserved sequences in phage genomes, which generate ambiguous cases that assemblers can not resolve with short reads alone. Several approaches can help to close the genomes such as PCR amplification based on the partially assembled genomes (Culley et al., 2007), or the use of a mix of long and short reads from the same sample as already done for microbial genomes (Boisvert, 2010).

The second major challenge of virus discovery through metagenomics is the extrapolation of characteristics beyond the genome sequence, with the most important one being the host range of the new virus. Except for the cases where a sequence identical (or nearly-identical) to the new virus is available in a sequenced microbial genome, assessing putative host is a tricky process. In the Far-T4 example, the putative host predicted by the different methods are all spurious, and non consistent. An *in silico* identification of putative host groups through genome composition (here tetranucleotide frequency) seems to be promising (Roux et al., in revision), and should be more and more efficient with the increasing coverage of microbial genome sequence space, as well as in cases where both a microbial and a viral metagenome are available from the same sample. However, such methodology will only provide a prediction of putative host that has to be verified by complimentary experiments like phageFISH (Allers et al., 2013), microfluidic digital PCR (Tadmor et al., 2011), or viral tagging (Deng et al., 2014) (reviewed in Dang and Sullivan, 2014; Brum and Sullivan, 2015). Assessing the impact of a virus also requires its quantification in different types of samples. If such quantification is now available for dsDNA viruses through linker-amplified metagenomes, there are no quantitative methodologies yet available for ssDNA and RNA viromes (Duhaime and Sullivan, 2012; Brum and Sullivan, 2015). Finally, the characterization of infection cycle through sequence analysis alone is hampered by the high number of "novelty" in each new viral genome (resulting in a lot of "hypothetical genes",) even for phages that are related to well-characterized isolates as the Far-T4 are from the T4 phages. Eventually, using these newly described genome sequences as anchors or probes for *in-situ* approaches

like phageFISH (Allers et al., 2013), meta-transcriptomics and viral meta-proteomics will be decisive to advance our knowledge of viral diversity beyond a first description of their genome and really characterize these new viruses.

Materials and Methods

Virome Generation and Assembly

The procedure used to generate viromes is the same as previously described (Roux et al., 2012). Briefly, each water sample was filtered on 0.22 μm, and virus-like particles were concentrated by tangential ultrafiltration and PEG precipitation (Colombet et al., 2007). These concentrates were treated with DNaseI to remove external fragments, before encapsidated DNA was freed via a thermal shock, purified with a QIAamp DNA mini kit (Qiagen), and randomly amplified with the phi29 polymerase using random hexamer primers (Genomiphi Kit, GE Healthcare). In a first study of two lakes (Lake Pavin and Lake Bourget), twenty liters of water were sampled at a 5 m depth in June and July 2008, and subjected (after preparation steps) to a single pyrosequencing run by GATC Biotech (Germany) using a 454 Life Sciences Genome Sequencer GS-FLX (Roux et al., 2012). Both virome read sets are available on the Short Read Archive (accession number: ERP000339). For this study, two samples were taken at 4 and 8 m on Lake Pavin in July 2013 and sequenced (after preparation) with Illumina HiSeq (GATC Biotech, Germany).

For 454 viromes, reads were first clustered using Uclust (Edgar, 2010) at a 100% identity level, in order to remove duplicate sequences, and sequence assembly was conducted with Newbler using threshold of 90% identity on at least 35 nucleotides. Illlumina sequences were trimmed by quality score (cutoff at 30 using FASTX v0.0.13) and then assembled using IDBA-UD (Peng et al., 2012).

Far-T4 Contig Selection

The major capsid protein Gp23, present in all T4 phages, is the only marker available for the Far-T4 group (Comeau and Krisch, 2008). First, all T4 sequences were identified by screening contigs for the presence of Gp23. Second, a phylogenetic tree of all virome sequences similar to Gp23 was computed in order to distinguish Far-T4 from other T4 phages and from putative false positive sequences. Thus, only Gp23 sequences found near the known Far-T4 sequences on the tree and displaying both N-terminal domain (coordinates 122–162) and C-terminal domain (coordinates 735–766) were kept (see Figure S5 for a complete view of the multiple alignment). For Illumina viromes, only genes detected on the contigs longer than 10 kb were selected to limit the total number of sequences in the analysis. All identified Far-T4 contigs were then automatically annotated with Metavir 2 (Roux et al., 2014b), which includes a gene prediction with Metagene Annotator (Noguchi et al., 2006), blastp comparison to NCBI Refseq genomes, and HMMER comparison with PFAM profiles (Punta et al., 2012), and are publicly available on Metavir (http://metavir-meb.univ-bpclermont.fr/) under project "FarT4 / Far-T4 Lake Pavin".

Sequence Analysis

Using proteins predicted from the Far-T4 contigs identified, phylogenies were inferred for different T4 phages conserved genes, namely the ones coding for the major capsid protein Gp23 (**Figure 1**), the portal protein Gp20, the large subunit of the terminase Gp17 and PhoH (Figure S1). For all these trees, reference sequences were obtained from the NCBI Refseq database of complete phage genomes, except for the g23 PCR amplicons that were obtained from the NCBI Genbank database. All phylogenies were based on (predicted) protein sequences.

Mutiple alignments were computed with Muscle (Edgar, 2004) and manually curated. The Gp23 multiple alignment was trimmed around the PCR amplicon boundaries to avoid artificially increased distances between sequences. FastTree2 (Price et al., 2010) was used to generate maximum-likelihood trees (WAG model). For all trees, all branches with bootstrap score lower than 50 were collapsed. The tree figures were edited with Itol (Letunic and Bork, 2007). The position of the root between the Far-T4 and all the other T4-like phages was determined by including an outgroup including *Spounavirinae* (another subfamily of *Myoviridae*).

Genome Fragment Comparison

To evaluate the "novelty" of Far-T4 contigs for each clade, the proportion of genes affiliated to NR, only similar to another Far-T4 contig or unique to a contig were calculated (Figure S2). A clustering of contigs was based on a blastp comparison of all vs. all predicted proteins from contigs. Genes were considered as shared when they displayed a blastp hit with a bit score greater than 50 and an *e*-value lower than 0.001. A proportion of shared genes between pairs of contigs was then computed as the number of genes shared between the two contigs divided by the length of the shortest contig. The cluster heatmap was computed in R with pheatmap package. For this analysis, duplicate contigs (*i.e.,* contigs 100% identical assembled from different samples) were excluded.

Finally, for the 12 contigs longer than 25 kb, the sequence comparison and map generation was performed using blastn (bit score >50) and Easyfig version 2.1 (Sullivan et al., 2011).

Major Capsid Protein Alignment and Structure Analysis

Jalview (Waterhouse et al., 2009) was used to display the multiple alignment of Gp23 as well as calculating residue conservation and consensus sequence. PaML (Yang, 2007) was used to calculate the dN/dS ratios and their associated likelihood value. Statistical tests were computed to detect the significance of likelihoods differences between evolutionary hypothesis as in (Zhang et al., 2005): (i) a single dN/dS ratio for all positions and all sequences, (ii) two dN/dS categories for all sequences, one linked to conserved sites, and one with sites under relaxed selection pressure, (iii) three different dN/dS for all sequences, one for conserved sites, one for relaxed selection sites, and one for sites under positive selections, and (iv) two different dN/dS for all sites, one for branches in the Far-T4 subtree, the other for all other branches (Table S1).

Secondary structures were predicted with I-Tasser (Roy et al., 2010) from the Gp23 sequence of contig Pavin_2013_4m-8. I-Tasser was also used to generate 3D models for representative sequences of each clade (based on the primary sequence contigs Pavin_2013_4m-8, Pavin_2013_4m-10, Pavin_2013_4m_77 and Pavin_2013_4m_62), based on the known structure of the conserved domain of the T4 vertex protein, which is shared with the T4 major capsid protein. For each major capsid protein, the stereochemical quality of each of the five models generated by I-Tasser for each sequence was assessed with ProSA-web (Wiederstein and Sippl, 2007), and the model with the best quality score on ProSA was kept. Model quality ranged from −5.4 to −7.56, in the range of X-Ray confirmed models for proteins of similar sizes (Figure S6). UCSF Chimera was used to display the different models as well as sequence conservation information (Pettersen et al., 2004).

Detection of Far-T4 Phages in Other Viral Metagenomes

Sequences similar to the large Far-T4 contigs assembled from Lake Pavin Illumina viromes, were searched in a large collection of viromes using tblastx (bit score >50, *e*-value < 0.001). Sequences similar to Far-T4 contigs were detected in seawater viromes from the Pacific Ocean Viromes (POV, Hurwitz and Sullivan, 2013), the Indian Ocean (Williamson et al., 2012), and the Atlantic Ocean (Chesapeake Bay, part of the GOS dataset Yooseph et al., 2007). Far-T4 sequences were also detected in several freshwater viromes from lakes in Europe (Roux et al., 2012), in Asia (Ge et al., 2013; Tseng et al., 2013), Antarctica (López-bueno et al., 2009), and in freshwater ponds in the USA (Rosario et al., 2009b). Conversely, no sequences similar to the Far-T4 were detected in other types of samples including human gut (Kim et al., 2011; Minot et al., 2012a), airborne samples (Whon et al., 2012) or plant samples (Coetzee et al., 2010).

Recruitment plots were generated with ggplot2 module in R, considering only BLAST hits with an amino-acid identity higher than 60% (in addition to bit score > 50 and *e*-value < 0.001). Coverage was calculated as the log10 of the number of reads mapped to the contig on sliding windows corresponding to a 30th of the contig length (i.e., for a contig of 30 kb, sliding windows of 1 kb would be used).

Host Prediction

Prophages or phages infecting single-cells (SAGs) closely related to Far-T4 were searched in microbial draft genomes by comparing predicted proteins from Far-T4 to the bacterial and archaeal genomes in Refseq and WGS NCBI database (blastp, bit score >50 and *e*-value < 0.001). In addition, CRISPR spacers were predicted on the bacterial and archaeal genomes available at Refseq and WGS (as of January 2014) with CRT (Bland et al., 2007) and then compared to the Far-T4 contigs with blastn. As CRISPR spacers are short sequences, more stringent thresholds were applied: only hits that covered more than 80% of the CRISPR spacer with more than 90% of nucleotide identity were considered significant. Three matches were identified: contigs Pavin_2013_4m_335 and Pavin_2013_4m_328 were similar to a CRISPR spacer from a *Streptococcus pneumoniae* genome (gi 452723578) at 92% of identity, and contig Pavin_2013_4m_8 matches a CRISPR spacer from another *Firmicutes, Peptoclostridium difficile* (gi 484228666), at 90% of identity.

Host prediction based on genomic signature was also computed using tetranucleotide frequency as in (Roux et al., in revision). First, tetranucleotide frequency vectors were calculated for each Far-T4 contig with Jellyfish (Marçais and Kingsford, 2011). The euclidean distance between these vectors and the tetranucleotide frequency vectors from bacterial and archaeal genomes in Refseq and WGS (as of January 2014) were then calculated. A previous analysis of more than 12,000 virus-host pairs indicated that in the absence of the exact host species in the database (which is the most likely case for freshwater viruses), host family could be predicted with 95% of success if the distance between virus and host tetranucleotide frequency vectors was below 4.10^{-04}, and with 84% of success if it was between 4.10^{-04} and 1.10^{-03}. For the Far-T4 phages, we could not detect any correspondence between Far-T4 contigs and microbial genomes displaying a distance lower than 2.10^{-04}, but three contigs from Clade 1 displayed a distance of $4{,}7.10^{-04}$ with genomes of *Ehrlichia chaffeensis* (two matching str. Arkansas–NC_007799.1, and one matching str. Sapulpa–GCF_000167655.1).

Acknowledgments

This work was performed under the auspices of the EC2CO program through the CAVIAR project led by FE, partially supported by a Gordon and Betty Moore Foundation grant (#3790) to MS, SR was partially supported by the University of Arizona Ecosystem Genomics Institute through a grant from the Technology and Research Initiative Fund through the Water, Environmental and Energy Solutions Initiative. An allocation of computer time from the UA Research Computing High Performance Computing (HPC) and High Throughput Computing (HTC) at the University of Arizona is gratefully acknowledged.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00199/abstract

References

Abrescia, N. G. A., Bamford, D. H., Grimes, J. M., and Stuart, D. I. (2012). Structure unifies the viral universe. *Annu. Rev. Biochem.* 81, 795–822. doi: 10.1146/annurev-biochem-060910-095130

Allen, L. Z., Ishoey, T., Novotny, M. A., McLean, J. S., Lasken, R. S., and Williamson, S. J. (2011). Single virus genomics: a new tool for virus discovery. *PLoS ONE* 6:e17722. doi: 10.1371/journal.pone.0017722

Allers, E., Moraru, C., Duhaime, M. B., Beneze, E., Solonenko, N., Canosa, J. B., 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. doi: 10.1111/1462-2920.12100

Bamford, D. H., Grimes, J. M., and Stuart, D. I. (2005). What does structure tell us about virus evolution? *Curr. Opin. Struct. Biol.* 15, 655–663. doi: 10.1016/j.sbi.2005.10.012

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., et al. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209. doi: 10.1186/1471-2105-8-209

Boisvert, S. (2010). Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J. Comput. Biol.* 17, 1519–1533. doi: 10.1089/cmb.2009.0238

Breitbart, M., Thompson, L. R., Suttle, C. A., and Sullivan, M. B. (2007). Exploring the vast diversity of Marine viruses. *Oceanography* 20, 135–139. doi: 10.5670/oceanog.2007.58

Brum, J., Culley, A., and Steward, G. (2013). Assembly of a Marine Viral Metagenome after physical fractionation. *PLoS ONE* 8:e60604. doi: 10.1371/journal.pone.0060604

Brum, J. R., Ignacio-espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., et al. (in press). Global patterns and ecological drivers of ocean viral communities. *Science*.

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. doi: 10.1038/nrmicro3404

Brüssow, H. (2009). The not so universal tree of life or the place of viruses in the living world. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 364, 2263–2274. doi: 10.1098/rstb.2009.0036

Butina, T. V., Belykh, O. I., Maksimenko, S. Y., and Belikov, S. I. (2010). Phylogenetic diversity of T4-like bacteriophages in Lake Baikal, East Siberia. *FEMS Microbiol. Lett.* 309, 122–129. doi: 10.1111/j.1574-6968.2010.02025.x

Charuvaka, A., and Rangwala, H. (2011). Evaluation of short read metagenomic assembly. *BMC Genomics* 12(Suppl 2), S8. doi: 10.1186/1471-2164-12-S2-S8

Chénard, C., and Suttle, C. A. (2008). Phylogenetic diversity of sequences of cyanophage photosynthetic gene psbA in marine and freshwaters. *Appl. Environ. Microbiol.* 74, 5317–5324. doi: 10.1128/AEM.02480-07

Coetzee, B., Freeborough, M.-J., Maree, H. J., Celton, J.-M., Rees, D. J. G., and Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology* 400, 157–163. doi: 10.1016/j.virol.2010.01.023

Colombet, J., Robin, A., Lavie, L., Bettarel, Y., Cauchie, H. M., and Sime-Ngando, T. (2007). Virioplankton "pegylation": use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *J. Microbiol. Methods* 71, 212–219. doi: 10.1016/j.mimet.2007.08.012

Comeau, A. M., and Krisch, H. M. (2008). The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol. Biol. Evol.* 25, 1321–1332. doi: 10.1093/molbev/msn080

Culley, A. I., Lang, A. S., and Suttle, C. A. (2007). The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. *Virol. J.* 4:69. doi: 10.1186/1743-422X-4-69

Dang, V. T., and Sullivan, M. B. (2014). Emerging methods to study viral infection at the single-cell level. *Front. Microbiol.* 5:724. doi: 10.3389/fmicb.2014.00724

Deng, L., Ignacio-Espinoza, J. C., Gregory, A., Poulos, B. T., Weitz, J. S., Hugenholtz, P., et al. (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* 513, 242–245. doi: 10.1038/nature13459

Dorigo, U., Jacquet, S., and Humbert, J.-F. (2004). Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. *Appl. Environ. Microbiol.* 70, 1017–1022. doi: 10.1128/AEM.70.2.1017

Duhaime, M. B., and Sullivan, M. B. (2012). Ocean viruses: rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology* 434, 181–186. doi: 10.1016/j.virol.2012.09.036

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 1–11. doi: 10.1038/ncomms5498

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

El Yacoubi, B., Bailly, M., and de Crécy-Lagard, V. (2012). Biosynthesis and function of posttranscriptional modifications of transfer RNAs. *Annu. Rev. Genet.* 46, 69–95. doi: 10.1146/annurev-genet-110711-155641

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B., and Banfield, J. F. (2012). Metagenomic assembly reveals dynamic viral populations in hypersaline systems. *Appl. Environ. Microbiol.* 78, 6309–6320. doi: 10.1128/AEM.01212-12

Filée, J., Tétart, F., Suttle, C. A., and Krisch, H. M. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12471–12476. doi: 10.1073/pnas.0503404102

Fuller, N., Wilson, W., Joint, I. R., and Mann, N. H. (1998). Occurrence of a sequence in marine cyanophages similar to that of T4 g20 and its application to PCR-based detection and quantification techniques. *Appl. Environ. Microbiol.* 64, 2051–2060.

Ge, X., Wu, Y., Wang, M., Wang, J., Wu, L., Yang, X., et al. (2013). Viral metagenomics analysis of Planktonic Viruses in East Lake, Wuhan, China. *Virologica Sinica*, 28, 280–290. doi: 10.1007/s12250-013-3365-y

Goldsmith, D. B., Crosti, G., Dwivedi, B., McDaniel, L. D., Varsani, A., Suttle, C., et al. (2011). Development of phoH as a novel signature gene for assessing marine phage diversity. *Applied and Environmental Microbiology*, 77, 7730–7739. doi: 10.1128/AEM.05531-11

Hall, A. R., Scanlan, P. D., Morgan, A. D., and Buckling, A. (2011). Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol. Lett.* 14, 635–642. doi: 10.1111/j.1461-0248.2011.01624.x

Hendrix, R. W. (2002). Bacteriophages: evolution of the majority. *Theor. Popul. Biol.* 61, 471–480. doi: 10.1006/tpbi.2002.1590

Hewson, I., Barbosa, J. G., Brown, J. M., Donelan, R. P., Eaglesham, J. B., Eggleston, E. M., et al. (2012). Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. *Appl. Environ. Microbiol.* 78, 6583–6591. doi: 10.1128/AEM.01705-12

Hjorleifsdottir, S., Aevarsson, A., Hreggvidsson, G. O., Fridjonsson, O. H., and Kristjansson, J. K. (2014). Isolation, growth and genome of the Rhodothermus RM378 thermophilic bacteriophage. *Extremophiles?*18, 261–270. doi: 10.1007/s00792-013-0613-x

Holmfeldt, K., Solonenko, N., Shah, M., Corrier, K., Riemann, L., VerBerkmoes, N. C., et al. (2013). Twelve previously unknown phage genera are ubiquitous in the global oceans. *Proc. Natl. Acad. Sci. U.S. A.* 110, 12798–12803. doi: 10.1073/pnas.1305956110

Hurwitz, B. L., Brum, J. R., and Sullivan, M. B. (2015). Depth-stratified functional and taxonomic niche specialization in the "core" and "flexible" Pacific Ocean Virome. *ISME J.* 9, 472–484. doi: 10.1038/ismej.2014.143

Hurwitz, B. L., Hallam, S. J., and Sullivan, M. B. (2013). Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* 14:R123. doi: 10.1186/gb-2013-14-11-r123

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. doi: 10.1371/journal.pone.0057355

Ignacio-Espinoza, J. C., and Sullivan, M. B. (2012). Phylogenomics of T4 cyanophages: lateral gene transfer in the "core" and origins of host genes. *Environ. Microbiol.* 14, 2113–2126. doi: 10.1111/j.1462-2920.2012.02704.x

Kang, I., Oh, H.-M., Kang, D., and Cho, J.-C. (2013). Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12343–12348. doi: 10.1073/pnas.1219930110

Kim, J. H., Son, J. S., Choi, Y. J., Choresca, C. H., Shin, S. P., Han, J. E., et al. (2012). Complete genome sequence and characterization of a broad-host range T4-like bacteriophage phiAS5 infecting Aeromonas salmonicida subsp. *salmonicida. Vet. Microbiol.* 157, 164–171. doi: 10.1016/j.vetmic.2011.12.016

Kim, M.-S., Park, E.-J., Roh, S. W., and Bae, J.-W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* 77, 8062–8070. doi: 10.1128/AEM.06331-11

Kim, S., Makino, K., Amemura, M., Shinagawa, H., and Nakata, A. (1993). Molecular analysis of the phoH gene, belonging to the phosphate regulon in *Escherichia coli*. *J. Bacteriol.* 175, 1316–1324.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23, 127–128. doi: 10.1093/bioinformatics/btl529

López-bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A., and Alcamí, A. (2009). High diversity of the Viral community from an Antarctic Lake. *Science* 326, 858–861. doi: 10.1126/science.1179287

Luo, C., Tsementzi, D., Kyrpides, N. C., and Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* 6, 898–901. doi: 10.1038/ismej.2011.147

Mann, N., Clokie, M., Millard, A., Cook, A., Wilson, W. H., Wheatley, P. J., et al. (2005). The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine Synechococcus strains. *J. Bacteriol.* 187, 3188–3200. doi: 10.1128/JB.187.9.3188

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Matteson, A. R., Loar, S. N., Bourbonniere, R. A., and Wilhelm, S. W. (2011). Molecular enumeration of an ecologically important cyanophage in a Laurentian Great Lake. *Appl. Environ. Microbiol.* 77, 6772–6779. doi: 10.1128/AEM.05879-11

Millard, A., Clokie, M. R. J., Shub, D. A., and Mann, N. H. (2004). Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. *Proc. Natl. Acad. Sci. U.S.A.* 101, 11007–11012. doi: 10.1073/pnas.0401478101

Miller, E. S., Heidelberg, J. F., Eisen, J. A., Nelson, W. C., Durkin, A. S., Ciecko, A., et al. (2003a). Complete genome sequence of the comparative genomics of a T4-Related Bacteriophage complete genome sequence of the Broad-Host-Range Vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriol.* 185, 5220–5233. doi: 10.1128/JB.185.17.5220

Miller, E. S., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T., and Rüger, W. (2003b). Bacteriophage T4 Genome. *Microbiol. Mol. Biol. Rev.* 67, 86–156. doi: 10.1128/MMBR.67.1.86

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012a). Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3962–3966. doi: 10.1073/pnas.1119061109

Minot, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012b). Conservation of Gene Cassettes among Diverse viruses of the Human Gut. *PLoS ONE* 7:e42342. doi: 10.1371/journal.pone.0042342

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28, 1420–1428. doi: 10.1093/bioinformatics/bts174

Petrov, V. M., Ratnayaka, S., Nolan, J. M., Miller, E. S., and Karam, J. D. (2010). Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol. J.* 7:292. doi: 10.1186/1743-422X-7-292

Pettersen, E., Goddard, T., and Huang, C. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Pride, D. T., Wassenaar, T. M., Ghose, C., and Blaser, M. J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics* 7:8. doi: 10.1186/1471-2164-7-8

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–301. doi: 10.1093/nar/gkr1065

Rappé, M. S., and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu. Rev. Microbiol.* 57, 369–394. doi: 10.1146/annurev.micro.57.030502.090759

Rohwer, F., Prangishvili, D., and Lindell, D. (2009). Roles of viruses in the environment. *Environ. Microbiol.* 11, 2771–2774. doi: 10.1111/j.1462-2920.2009.02101.x

Rosario, K., Duffy, S., and Breitbart, M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J. Gen. Virol.* 90(Pt 10), 2418–2424. doi: 10.1099/vir.0.012955-0

Rosario, K., Nilsson, C., Lim, Y. W., Ruan, Y., and Breitbart, M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environ. Microbiol.* 11, 2806–2820. doi: 10.1111/j.1462-2920.2009.01964.x

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the Diversity and specificity of two freshwater Viral Communities through Metagenomics. *PLoS ONE* 7:e33641. doi: 10.1371/journal.pone.0033641

Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., et al. (2014a). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta- genomics. *eLife* 3, 1–20. doi: 10.7554/eLife.03125

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014b). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15, 1–12. doi: 10.1186/1471-2105-15-76

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Sabri, M., Häuser, R., Ouellette, M., Liu, J., Dehbi, M., Moeck, G., et al. (2011). Genome annotation and intraviral interactome for the Streptococcus pneumoniae virulent phage Dp-1. *J. Bacteriol.* 193, 551–562. doi: 10.1128/JB.01117-10

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D. B., et al. (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J.* 1, 492–501. doi: 10.1038/ismej.2007.67

Sullivan, M. B. (2015). Viromes, not gene markers for studying dsDNA viral communities. *J. Virol.* 89, 2459–2461. doi: 10.1128/JVI.03289-14

Sullivan, M. B., Coleman, M. L., Quinlivan, V., Rosenkrantz, J. E., Defrancesco, A. S., Tan, G., et al. (2008). Portal protein diversity and phage ecology. *Environ. Microbiol.* 10, 2810–2823. doi: 10.1111/j.1462-2920.2008.01702.x

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* 3:e144. doi: 10.1371/journal.pbio.0030144

Sullivan, M. B., Huang, K. H., Ignacio-Espinoza, J. C., Berlin, A. M., Kelly, L., Weigele, P. R., et al. (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* 12, 3035–3056. doi: 10.1111/j.1462-2920.2010.02280.x

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., and Chisholm, S. W. (2006). Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and their hosts. *PLoS Biol.* 4:e234. doi: 10.1371/journal.pbio.0040234

Sullivan, M. J., Petty, N. K., and Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics* 27, 1009–1010. doi: 10.1093/bioinformatics/btr039

Suttle, C. A. (2007). Marine viruses–major players in the global ecosystem. *Nature Reviews Microbiology* 5, 801–812. doi: 10.1038/nrmicro1750

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. doi: 10.1126/science.1200758

Tseng, C.-H., Chiang, P.-W., Shiah, F.-K., Chen, Y.-L., Liou, J.-R., Hsu, T.-C., et al. (2013). Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J.* 7, 2374–2386. doi: 10.1038/ismej.2013.118

Vázquez-Castellanos, J. F., García-López, R., Pérez-Brocal, V., Pignatelli, M., and Moya, A. (2014). Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 15:37. doi: 10.1186/1471-2164-15-37

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Weigele, P. R., Pope, W. H., Pedulla, M. L., Houtz, J. M., Smith, A. L., Conway, J. F., et al. (2007). Genomic and structural analysis of Syn9, a cyanophage infecting marine Prochlorococcus and Synechococcus. *Environ. Microbiol.* 9, 1675–1695. doi: 10.1111/j.1462-2920.2007.01285.x

Whon, T. W., Kim, M.-S., Roh, S. W., Shin, N.-R., Lee, H.-W., and Bae, J.-W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J. Virol.* 86, 8221–8331. doi: 10.1128/JVI.00293-12

Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* 35, W407–W410. doi: 10.1093/nar/gkm290

Williamson, S. J., Allen, L. Z., Lorenzi, H. A., Fadrosh, D. W., Brami, D., Thiagarajan, M., et al. (2012). Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS ONE* 7:e42047. doi: 10.1371/journal.pone.0042047

Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., et al. (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* 3:e1456. doi: 10.1371/journal.pone.0001456

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5:e16. doi: 10.1371/journal.pbio.0050016

Zeidner, G., Bielawski, J. P., Shmoish, M., Scanlan, D. J., Sabehi, G., Béjà, O., et al. (2005). Potential photosynthesis gene recombination between. *Environ. Microbiol.* 7, 1505–1513. doi: 10.1111/j.1462-2920.2005.00833.x

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. doi: 10.1093/molbev/msi237

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., et al. (2013). Abundant SAR11 viruses in the ocean. *Nature* 494, 357–360. doi: 10.1038/nature11921

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

frontiers in Microbiology

ORIGINAL RESEARCH
published: 08 November 2016
doi: 10.3389/fmicb.2016.01740

Ultrastructure and Viral Metagenome of Bacteriophages from an Anaerobic Methane Oxidizing *Methylomirabilis* Bioreactor Enrichment Culture

Lavinia Gambelli[1], *Geert Cremers*[1], *Rob Mesman*[1], *Simon Guerrero*[1], *Bas E. Dutilh*[2,3], *Mike S. M. Jetten*[1], *Huub J. M. Op den Camp*[1] *and Laura van Niftrik*[1]*

[1] *Department of Microbiology, Faculty of Science, Institute for Water and Wetland Research, Radboud University, Nijmegen, Netherlands,* [2] *Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands,* [3] *Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, Netherlands*

OPEN ACCESS

Edited by:
William Michael McShan, University of Oklahoma Health Sciences Center, USA

Reviewed by:
Bernard A. P. Lafont, National Institute of Allergy and Infectious Diseases, USA
Claire Bertelli, Simon Fraser University, Canada

***Correspondence:**
Laura van Niftrik
l.vanniftrik@science.ru.nl

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *07 July 2016*
Accepted: *18 October 2016*
Published: *08 November 2016*

Citation:
Gambelli L, Cremers G, Mesman R, Guerrero S, Dutilh BE, Jetten MSM, Op den Camp HJM and van Niftrik L (2016) Ultrastructure and Viral Metagenome of Bacteriophages from an Anaerobic Methane Oxidizing Methylomirabilis Bioreactor Enrichment Culture. Front. Microbiol. 7:1740. doi: 10.3389/fmicb.2016.01740

With its capacity for anaerobic methane oxidation and denitrification, the bacterium *Methylomirabilis oxyfera* plays an important role in natural ecosystems. Its unique physiology can be exploited for more sustainable wastewater treatment technologies. However, operational stability of full-scale bioreactors can experience setbacks due to, for example, bacteriophage blooms. By shaping microbial communities through mortality, horizontal gene transfer, and metabolic reprogramming, bacteriophages are important players in most ecosystems. Here, we analyzed an infected *Methylomirabilis* sp. bioreactor enrichment culture using (advanced) electron microscopy, viral metagenomics and bioinformatics. Electron micrographs revealed four different viral morphotypes, one of which was observed to infect *Methylomirabilis* cells. The infected cells contained densely packed ∼55 nm icosahedral bacteriophage particles with a putative internal membrane. Various stages of virion assembly were observed. Moreover, during the bacteriophage replication, the host cytoplasmic membrane appeared extremely patchy, which suggests that the bacteriophages may use host bacterial lipids to build their own putative internal membrane. The viral metagenome contained 1.87 million base pairs of assembled viral sequences, from which five putative complete viral genomes were assembled and manually annotated. Using bioinformatics analyses, we could not identify which viral genome belonged to the *Methylomirabilis*- infecting bacteriophage, in part because the obtained viral genome sequences were novel and unique to this reactor system. Taken together these results show that new bacteriophages can be detected in anaerobic cultivation systems and that the effect of bacteriophages on the microbial community in these systems is a topic for further study.

Keywords: *Methylomirabilis*, bacteriophage, viral metagenome, ultrastructure, bioreactor

INTRODUCTION

The importance of microorganisms in biogeochemical cycles and global warming is well-known (Falkowski et al., 2008). Particularly, microorganisms that take part in the carbon and nitrogen cycles are of great interest for both the scientific community and the public at large due to the growing awareness of climate change. In fact, among greenhouse gasses that affect the climate, carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) are the most relevant (Pachauri et al., 2014).

Aerobic and anaerobic methanotrophic microorganisms oxidize CH_4 to CO_2 mitigating the release of CH_4 to the atmosphere. The newly discovered anaerobic denitrifying methanotroph "*Candidatus Methylomirabilis oxyfera*" was enriched from freshwater sediments and couples methane oxidation to nitrite reduction, thereby linking carbon and nitrogen cycles via a postulated unique intra-aerobic pathway (Ettwig et al., 2009, 2010). Next to its remarkable physiology, *M. oxyfera* stands out from other bacteria also with respect to its cell shape and structure. These Gram-negative, rod-shaped bacteria have a polygonal cell shape constituted by several longitudinal ridges running along the cell length and converging in a cap-like structure at the cell poles (Wu et al., 2012). Furthermore, these microorganisms are not only interesting from a fundamental scientific point of view, but could also be important players in natural ecosystems (Deutzmann et al., 2014; Hu et al., 2014) and be implemented in the removal of dissolved methane and ammonium from digester effluents in combination with anaerobic ammonium oxidizing bacteria (Luesken et al., 2011; Shi et al., 2013).

Bacteria constitute the vast majority of the biomass on Earth (about 90%), yet viruses are the most abundant "biological entities," comprising ∼94% of the nucleic-acid containing-particles (Suttle, 2007). The discovery that viruses indiscriminately occupy marine environments (Suttle, 2005) and freshwater ecosystems (Middelboe et al., 2008) opened a new research area concerning the effect of bacteriophages on microbial populations and the role of bacteriophages in elemental and nutrient cycles. By inducing bacterial lysis, bacteriophages affect abundance, diversity and functioning of microbial populations (Weinbauer and Rassoulzadegan, 2004).

Marine and freshwater viruses show a great variety of different morphologies (Wommack and Colwell, 2000; Sulcius et al., 2011; Borrel et al., 2012). Viruses have a huge heterogeneity in size and genetic structure. The size ranges from 20 to 200 nm in diameter (Brum et al., 2013), and the genetic material can be either dsDNA, ssDNA (+), dsRNA, or ssRNA (±) (Dimmock et al., 2007). Viruses infecting Bacteria and Archaea generally consist of a proteinaceous capsid that contains the genetic material. They can also have a tail, which is used to inject the genetic material into the host. Tailed viruses belong to three families: the *Siphoviridae* (non-contractile, long tail), the *Myoviridae* (contractile, medium-length tail), and the *Podoviridae* (non-contractile, short tail). Tailless viruses have been described with very diverse morphologies, such as polyhedral, filamentous, spindle-like, cubic-like, star-like, or pleomorphic. Among tailless viruses, bacteriophages with an internal lipid bilayer enclosing the genome have been described and assigned to the *Tectiviridae* and *Corticoviridae* families (King et al., 2012).

The striking morphological diversity of viruses is also reflected in their great genetic heterogeneity (Breitbart et al., 2002). Bacteriophages assigned to a certain family often share very little or no sequence similarity. Comparative analysis of bacteriophage genomes often reveals that genes are organized in modules characterized by a different evolutionary history. This phenomenon is known as genome mosaicism, and it is the result of a high degree of horizontal gene exchange (Hatfull, 2008).

The lack of a universal genetic marker and the fact that only a minority of bacterial hosts can be cultivated in laboratories (Edwards and Rohwer, 2005) make the assessment of such heterogeneity very challenging. Although, culture-independent, large-scale metagenomics has paved the way to a more exhaustive understanding of the viral sequence space (Dutilh, 2014), this approach is not flawless. For example, small datasets may be obtained due to the limited yield of viral DNA from environmental samples, biasing downstream analysis. Moreover, viruses show a high spatiotemporal variability (Koskella and Parr, 2015), contributing to a limited overlap between samples. The analysis of viral metagenomes is laborious, partly because of the lack of reference databases and appropriate analytical tools compared to, for example, the ones for microbial metagenomics (Mokili et al., 2012; Hurwitz and Sullivan, 2013). Consequently, a high abundance of unknown viral sequences (65–95%) are reported in new surveys when these metagenomes are mapped against databases of unknown sequences (Breitbart et al., 2007; Mokili et al., 2012). Finally, viral metagenomics studies are frequently still relatively small in scope.

Several studies investigated the complex dynamics of phage-host interactions in lab-scale bioreactor systems (for example Barr et al., 2010; Shapiro et al., 2010). However, there are limited bioreactor studies on virus-host interaction that include viral metagenomics (one example is Kunin et al., 2008). The present paper describes the bacteriophage population in a *Methylomirabilis* bioreactor enrichment culture using (advanced) electron microscopy, viral metagenomics, and bioinformatics analysis. Several viral genomes and morphologies were found and one lytic bacteriophage was observed to infect *Methylomirabilis* cells. Through bioinformatics we attempted to find which of the obtained viral contigs belonged to the bacteriophage that infects *Methylomirabilis* cells. Finally we speculate on the significance of the infection on bacterial growth and population dynamics.

MATERIALS AND METHODS

Enrichment Conditions

The *Methylomirabilis* enrichment culture (∼80% *Methylomirabilis* sp.) was grown in a continuous sequencing batch reactor (Applicon Biotechnology BV, Applisens, Schiedam, the Netherlands) made of stainless steel and glass with a volume of 6 L. The inoculum biomass was derived from a pre-existing enrichment, originally inoculated with sediment samples from a ditch in the Ooijpolder (Ettwig et al., 2009). The culture was kept anoxic by a continuous supply of a gas mixture composed of methane and carbon dioxide (95:5, v/v) and the medium was

continuously flushed with a mixture of argon and carbon dioxide (95:5, v/v). The pH was maintained at 7.3 ± 0.1, the temperature was kept stable at 30°C and the system was constantly stirred at 100 RPM. The volume of the enrichment was kept at 4 L by a level sensor controlled pump in sequential cycles of feeding (22.5 h) and rest (30 min to settle, 60 min to pump out excess medium). The HRT (hydraulic retention time) was 4 days and the composition of the medium was (per L) 0.552–2.07 g $NaNO_2$ (8–30 mM depending on culture performance), 288 μg $MgSO_4$, 192 μg $CaCl_2$, 50 μg KH_2PO_4, 2.5 mg $FeSO_4{*}7H_2O$. The trace element solution was adapted from Ettwig et al. (2010) and contained per L: 150 μg $ZnSO_4.7H_2O$, 60 μg $CoCl_26H_2O$, 400 μg $CuSO_4$., 100μg $NiCl_26H_2O$, 10 μg H_3BO_3, 100 μg $MnCl_24H_2O$, 10 μg $Na_2WO_4.2H_2O$, 50 μg $Na_2MoO_4.2H_2O$, 15 μg SeO_2, 10 μg $CeCl_2$.

Negative Staining of Bacteriophages Isolated from the *Methylomirabilis* Enrichment Culture

A 20 ml sample was collected from the bioreactor enrichment culture and filtered with 0.45 and 0.2 μm syringe filters, sequentially (Puradisc Cellulose Acetate, Whatman) to remove microorganisms and keep only the viral fraction. The flow-through was concentrated using spin columns (Vivaspin, GE Healthcare, 10,000 MWCO) at 5000 × g for 5 min at 4°C, (Allegra X-15R Centrifuge, Beckman Coulter) to a final volume of about 60 μl. Approximately 3 μl sample was placed on glow-discharged and carbon-Formvar-coated 100 mesh hexagonal copper grids (Veco) and incubated at room temperature for 15 min. The excess liquid was blotted off using filter paper (Ashless, Grade 589/1 Filtration Paper, Whatman). The specimens were stained with 0.5% uranyl acetate for 1 min, blotted dry, washed in Milli-Q, blotted dry with filter paper and let air dry completely. The virioplankton was observed using a JEOL JEM-2100 transmission electron microscope operated at 200 kV.

Freeze-Etching on the *Methylomirabilis* Enrichment Culture

A 2 ml sample was collected from the bioreactor enrichment culture and centrifuged at 800 × g for 4 min at 30°C (Microcentrifuge 5417R, Eppendorf, Hamburg, Germany). Supernatant was decanted and the pellet was resuspended in remaining supernatant. For high-pressure freezing, a specimen sandwich was assembled in the specimen holder of a Leica EMHPF. The sandwich consisted of an aluminum spacer ring (0.2 × 3 mm) and a membrane carrier (2.6 × 0.1 mm, no. 16707898) in which 0.2 μl sample was loaded, the sample was sealed off with a gold stub (specimen carrier D3/D2, gold, no. 16770131). After closing the holder, the sample was high-pressure frozen in an EMHPF, operating at 2100 bar. Samples were stored in liquid nitrogen.

To obtain a freeze-fracture replica, the sample was placed in a detachable cold table and loaded onto the stage of a Balzers BAF400 freeze-etch machine, pre-cooled to −170°C. Specimens were fractured at −110°C and subsequently etched for 12 min at the same temperature with a vacuum below 10^{-7} Bar. The specimen was shadowed with 1.3 nm Pt (angle 45°) and 15 nm C (angle 90°) and a replica of the sample was obtained. The biological material underneath the replica was digested in 70% H_2SO_4 for 48 h. Replicas were washed twice in MilliQ water and picked up with glow discharged 700 mesh hexagonal copper grids (reference number G276OG, Agar Scientific) and investigated by a JEOL JEM-2100 transmission electron microscope operated at 200 kV.

Cryofixation, Freeze-Substitution, Epon Embedding, Thin-Sectioning, and Post-Staining of *Methylomirabilis* Enrichment Culture

Cells from the bioreactor enrichment culture were harvested and cryofixed by high-pressure freezing (Leica HPM 100; Leica Microsystems, Vienna, Austria). Samples were placed into a 100 μm cavity of a type A platelet (3 mm diameter; 0.1–0.2 mm depth, Leica Microsystems) and closed with the flat side of a lecithin-coated type B platelet (3 mm diameter, 0.3 mm depth). The platelets were stored in liquid nitrogen.

For epon embedding, frozen samples were freeze-substituted in acetone containing 2% OsO_4, 0.2% uranyl acetate, and 1% H_2O (Walther and Ziegler, 2002). The substitution followed several intervals: cells were kept at −90°C for 47 h; brought to −60°C at 2°C per hour and kept at −60°C for 8 h; brought to −30°C at 2°C per hour and kept at −30°C for 8 h in a freeze-substitution unit (AFS; Leica Microsystems, Vienna, Austria). To remove uranyl acetate, the samples were washed four times for 30 min in the AFS device at −30°C with acetone containing 2% $OsO_4$and 1% H_2O. Next, fixation was continued on ice for 1 h. OsO_4 and H_2O were removed by washing the sample twice in anhydrous acetone for 30 min. Samples were gradually infiltrated with epon resin and polymerized for 72 h at 60°C (Mollenhauer, 1964). Ultrathin sections of 60 nm were cut using a Leica UCT ultramicrotome (Leica Microsystems, Vienna, Austria) and collected on carbon-Formvar-coated 100 mesh hexagonal copper grids (Veco). The sections were then post-stained with 2% uranyl acetate for 20 min and lead citrate for 2 min. After each of the two steps, grids were washed in MilliQ water. Sections were investigated using a JEOL JEM1010 transmission electron microscope operated at 60 kV.

Electron Tomography on Infected *Methylomirabilis* Cells

Semi-thin sections (200–300 μm) were cut using a diamond knife (Diatome, Biel, Switzerland) and an ultramicrotome (UCT, Leica microsystems) and collected on 50 mesh copper grids with a carbon coated formvar support film. After air drying, grids were post-stained with 4% uranyl acetate in MilliQ for 30 min and Reynolds lead citrate for 2 min. Ten nanometers proteinA gold (CMC, UMC Utrecht, The Netherlands) was applied to the sections to act as fiducial marker during tilt-series acquisition and reconstruction.

Virus infected cells and isolated virus particles were chosen as the region of interest. Dual axis tilt series (−60 to +60) were recorded on the JEOL JEM-2100 microscope operating at 200 kV,

using SerialEM for automated image acquisition (Mastronarde, 2005).

Recorded tilt series were reconstructed using the IMOD package (Kremer et al., 1996) and tomograms were generated using both the weighted back-projection and SIRT algorithms. Reconstructed tomograms were segmented by hand using 3DMOD. Iso-surface model was generated from a sub-tomogram employing a circular mask. Summed virtual slices (36 slices) were visualized using the UCSF Chimera package (Pettersen et al., 2004), the histogram was adjusted to fit the density associated with the virus particle.

Concentration of Viral Particles by Iron Chloride and PEG 8000 Precipitation

Over a period of ∼3 months bioreactor material and effluent (13 L) were collected from the bioreactor enrichment culture and stored at 4°C until further analyses. Since the majority of the microbial population in the bioreactor grows in aggregates, the total sample was split in two: the bioreactor supernatant containing free bacteriophages and the bacterial biomass containing bacteriophages in the aggregates and within the cells. The two samples were obtained by centrifuging the sample at 20,000 × g for 1 h in 350 ml centrifuge tubes (Thermo Scientific, SorvallTM LYNX 4000 centrifuge). The resulting biomass pellets were resuspended in a total volume of 50 mL of the supernatant. Free-floating bacteriophages were precipitated by iron chloride flocculation, whereas bacterial biomass-associated bacteriophages were collected by PEG 8000.

Iron Chloride Flocculation

To isolate free bacteriophages approximately 12.95 L of supernatant was first filtered with 0.45 and 0.2 μm filters sequentially (nuclepore track-etched polycarbonate membrane filters, Whatman) to remove microbial biomass. The sample was subsequently processed by iron chloride flocculation as described previously (John et al., 2011; Cunningham et al., 2015). Since iron is known to be an inhibitor of DNA amplification, it was removed from the sample by extensively washing the sample with MSM buffer (400 mM NaCl, 20 mM MgSO4·Ů7H2O, 50 mM Tris, pH 7.5) in spin columns (Vivaspin, GE Healthcare, 10,000 MWCO) at 2500 × g for 40 min at 4°C, using an Allegra X-15R Centrifuge (Beckman Coulter). The sample was stored at 4°C until further analyses.

PEG 8000 Precipitation

About 50 ml of microbial biomass was first disrupted by pottering on ice to free the bacteriophages. To digest non-viral DNA and RNA, DNase I (Thermo Scientific, final concentration 1U/ml), and RNase A (Thermo Scientific, final concentration 5U/mL) treatments were simultaneously performed for 1 h at 37°C. The sample volume was divided in 2 ml Eppendorf tubes and PEG 8000 precipitation was performed as described previously (Guo et al., 2012). The sample was centrifuged for 30 min at 2000 × g and the obtained pellet was resuspended in MSM buffer. PEG 8000 was separated from the viral particles as described previously (Colombet and Sime-Ngando, 2012). In addition, to lower the salt concentration, the supernatant containing free phages was extensively washed with MSM buffer in spin columns (Vivaspin, GE Healthcare, 10,000 MWCO) at 2500 × g and 4°C using an Allegra X-15R Centrifuge (Beckman Coulter) to a final concentration of 20 mM KCl. The sample was stored at 4°C until further analyses.

DNA Extraction

The two viral samples obtained by iron chloride and PEG 8000 precipitation were pooled together and bacteriophages were further concentrated by ultracentrifugation (Optima XE90, Beckman-Coulter, rotor type 90 Ti, Beckman-Coulter) at 77,000 rcf for 1 h at 4°C (Szpara et al., 2011). A P1 reference bacteriophage NC_005856.1 was used as a positive control for DNA extraction. The pellet was resuspended in 1 ml of supernatant and the total DNA was extracted according to the protocol published by Thurber et al. (2009). Using the Qubit dsDNA HS assay kit (Life), the extracted DNA was quantified at 0.2 ng DNA.

Library Preparation and Sequencing

To prevent amplification biases, yet obtain enough DNA for Ion Torrent sequencing, 43.2 ng of 16S ribosomal DNA of "*Candidatus* K. stuttgartiensis" was added to 0.1 ng of viral DNA (Cremers et al., in preparation). The total DNA was sheared for 6 cycles using the Bioruptor (Diagenode, Liege, Belgium; 1 min on, 1 min off) and prepared according to manufactures protocol (IonXpress Plus gDNA fragment library preparation Rev C.0, Life). The sample was sequenced using an Ion Torrent Personal Genome Machine (Life) on a v318 chip following manufacturer's protocol.

Bioinformatics

Contig Assembly

The sequencing reads (4,334,460) were trimmed with default quality settings (size 25–375 bp) using CLC genomics workbench (CLCbio, Aarhus, Denmark) and filtered to remove 16S ribosomal *K. stuttgartiensis* DNA (∼87.6%) and *M. oxyfera* genomic DNA (∼0.08%). The remaining reads were assembled using SPAdes (v.3.5.0), and ESOM (Ultsch and Mörchen, 2005; Dick et al., 2009) was used to validate the consistency of the contig tetranucleotide profiles (Supplementary File 1). This resulted in 2094 assembled contigs with a total length of 1.87 Mbp, including the five proposed genome sequences (Supplementary Files 2A–C).

Annotation of Genes on the Bacteriophage Contigs

Putative ORFs (open reading frames) were predicted and automatically annotated on all contigs (Supplementary File 3) by using Prokka (Seemann, 2014). The five longest and putative complete viral genomes were selected for manual annotation, and visualized by ARTEMIS (Rutherford et al., 2000). The automatic annotation of each putative ORF was manually curated using BLASTp (Johnson et al., 2008). An $e < 10^{-3}$ was chosen as a cut-off for annotation of genes and domains. The amino acid sequences were also analyzed for the presence of conserved domains consulting: InterProScan (Jones et al., 2014), PROSITE (De Castro et al, 2006), Motif Scan (Pagni

et al., 2007), Pfam (Finn et al., 2016), and SMART (Letunic et al., 2014). For each putative ORF, percentage identity and overlap to the first annotated/validated hit obtained through BLASTp were calculated using Pairwise Alignment (Wu et al., 2003). Classification of annotated genes was performed according to the following criteria. Hypothetical protein; >20% identity to hypothetical protein, conserved hypothetical protein; >30% identity to hypothetical protein (similar length), similar to; >20% identity to known protein (similar length), strongly similar to; >40% identity to known protein (similar length). ORFs that contained conserved domains (but no annotated Blast hits) or (strong) similarity to a known protein but different length were described as putative [protein name].

To assess the presence of known viruses and bacteria in the viral metagenome, all contigs obtained with SPAdes and their predicted ORFs were compared against the nr databases from NCBI (using BLASTn and BLASTp, respectively) and the Earth's virome database (Paez-Espino et al., 2016; Supplementary File 4).

Accession Numbers

The viral genomes are available from GenBank under accession numbers, KX853510, KX853511, KX853512, KX853513, KX853514 for Genomes 1–5, respectively. The remaining contigs are submitted under number MKFH00000000.

Predicting Which of the Sequenced Bacteriophages Infects *Methylomirabilis* Sp.

Computational phage-host prediction signals (Edwards et al., 2015) were calculated to identify which of the 2094 assembled contigs belonged to the bacteriophage infecting *Methylomirabilis*. First, genetic homology between the *Methylomirabilis* genomes (*Methylomirabilis oxyfera*; Ettwig et al., 2010 and a new *Methylomirabilis* species; Guerrero et al., in preparation) and the viral contigs was determined by using blastn and tblastx 2.2.30+ with default settings and $E \leq 10^{-5}$ (Camacho et al., 2009), and the total bitscore for each viral contig was recorded. Second, both *Methylomirabilis* genomes were checked for CRISPR sequences using CRISPRfinder (http://crispr.u-psud.fr/Server/) and CRASS (Skennerton et al., 2013). The resulting spacer sequences were subsequently cross-referenced with the viral genomes and contigs from the SPAdes assembly using CRISPRtarget (Biswas et al., 2013). Third, similarity in oligonucleotide usage was calculated as 1—E, where E is the Euclidean distance between the k-mer profiles of each viral contig and the *Methylomirabilis* genomes ($k = 2, 4, 6$).

RESULTS

Bacteriophage Population in a *Methylomirabilis* Bioreactor Enrichment Culture

We investigated the bacteriophage population in a bioreactor containing an enrichment culture of the anaerobic methanotroph *Methylomirabilis* sp. (**Table 1**). Several free bacteriophages with different morphologies were observed using negative staining. Among these, four morphotypes were identified. All bacteriophages had putative icosahedral capsid symmetry and three of them possessed a tail. Capsids and tails varied in diameter and length, thereby allowing a clear distinction between the morphotypes (**Figure 1**).

Bacteriophage Infection in *Methylomirabilis* Cells

In addition to free bacteriophages, the microbial population was investigated for phage infection using thin sections, freeze-etching, and electron tomography. Thin sections of high-pressure frozen, freeze-substituted, and epon embedded biomass samples showed a lytic phage infection in *Methylomirabilis* cells (**Figure 2**). The infection rate (amount of infected *Methylomirabilis* cells in the total *Methylomirabilis* population) was calculated to be 2.3%. The infection did not affect bioreactor performance with respect to activity (methane and nitrite consumption). The intracellular bacteriophages had a hexagonal capsid (ca. 55 nm wide, vertex-to-vertex), indicating an icosahedral 3D symmetry. The capsid contained an electron dense and round central core (ca. 20 nm diameter), probably enclosing the genetic material. The bacteriophages were present both outside and inside the *Methylomirabilis* cells. Infected *Methylomirabilis* cells displayed different stages of bacteriophage assembly. In addition to completely assembled bacteriophages, entities were observed inside the cells that contained only the central core, i.e., the capsid was not yet assembled. As more and more bacteriophages were assembled, they were organized in a highly packed formation within the cell (**Figures 2C,D**). This caused the infected cell to swell approximately 1.9x in size compared to a not infected cell [area measured on TEM sections based on 10 infected ($0.44 \pm 0.2\ \mu m^2$) and 10 not infected cells ($0.23 \pm 0.06\ \mu m^2$)]. Afterwards the infected

TABLE 1 | Properties of the 4 viral morphotypes observed in the *Methylomirabilis* enrichment culture.

Bacteriophage	Capsid size (nm)*	Capsid symmetry	Tail length (nm)*	Tail width (nm)*	Morphotype
Figure 1A	~66	Icosahedral	–	–	Non tailed virus
Figure 1B	~87	Putative icosahedral	~90	~21	Myoviridae-like
Figure 1C	~50	Putative icosahedral	~87	~8.5	Myoviridae-like
Figure 1D	~41	Putative icosahedral	~91	~16	Myoviridae-like

**All reported measures were obtained from negative-stained samples.*

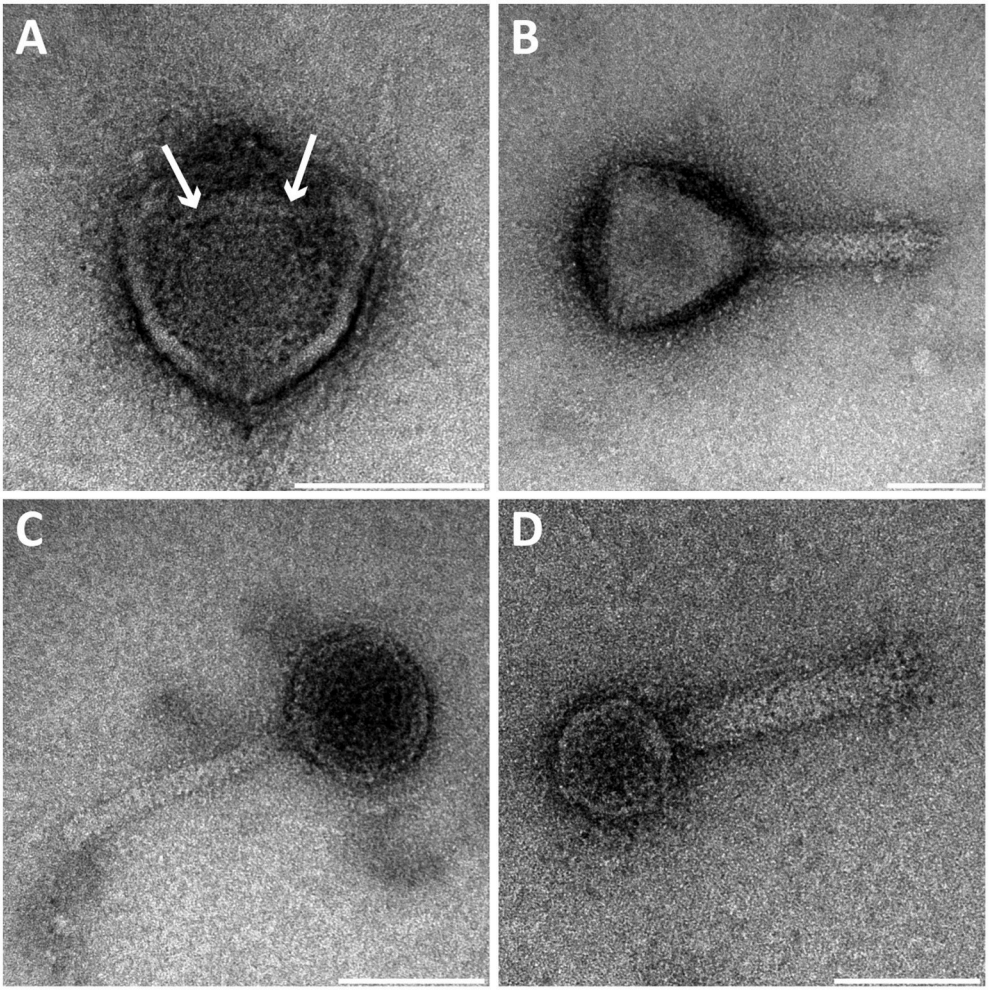


FIGURE 1 | Transmission electron micrographs of negative-stained bacteriophages present in the *Methylomirabilis* bioreactor enrichment culture. Four viral morphotypes were observed. **(A)** Viral morphotype with putative icosahedral symmetry of the head (~71 nm diameter), without a tail or other appendages but with a central round internal core (~25 nm diameter, see arrows). **(B–D)** Three additional viral morphotypes all with a tail and a putative icosahedral symmetry of the head (~87, ~50, and ~41 nm diameter respectively). Scale bars; 50 nm.

cell lysed thereby releasing the bacteriophages (**Figures 2E,F**). Comparing the intracellular *Methylomirabilis* bacteriophage to the free bacteriophages observed in the bioreactor enrichment culture using negative staining, its morphology resembled that of the bacteriophage depicted in **Figure 1A** based on size and symmetry. Freeze-etching of infected *Methylomirabilis* cells revealed the surface of the bacteriophage capsid (**Figure 3**). The capsomeres, the building subunits of the capsid, were arranged in triangular faces that constitute the icosahedral symmetry of the proteinaceous capsid.

Electron tomography was used to study the 3D structure of the bacteriophage and the infected *Methylomirabilis* cells (**Figure 4**). 3D reconstruction and modeling showed that the cytoplasmic membrane of infected cells was broken at many places while the cell wall was often still intact (**Figures 4A,B** and Supplementary Movie 1). Inside infected cells, both completely assembled bacteriophages were present as well as incompletely assembled bacteriophages (with only the electron dense core visible—no capsid yet). 3D reconstruction and modeling of free bacteriophages suggested the presence of a putative internal membrane surrounding the electron dense core (**Figures 4C–F**). In addition, no tail was observed on the free bacteriophages.

Viral Metagenome of the Bacteriophage Population in the *Methylomirabilis* Bioreactor Enrichment Culture

To characterize the bacteriophage population present in the *Methylomirabilis* bioreactor enrichment culture, the total viral DNA was extracted from the bioreactor and the effluent and sequenced. After assembly, 2094 contigs were obtained (Supplementary Files 2A,B). The five longest and putative complete viral genomes (**Table 2**) were selected for manual annotation (**Figure 5** and Supplementary Files 5A–E). Four of these contigs (197, 85, 71, 16 kb) contained identical sequences

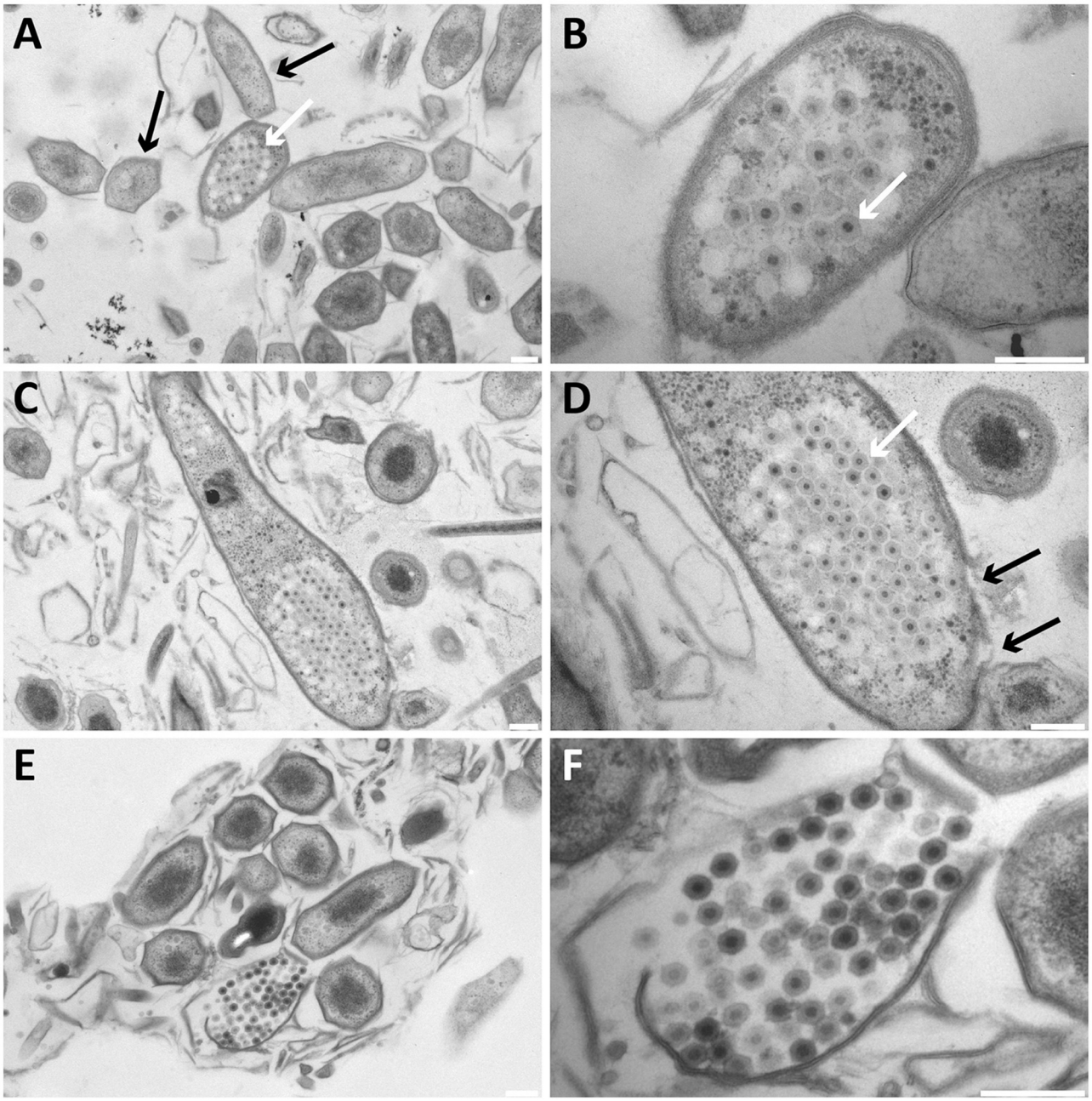


FIGURE 2 | Transmission electron micrographs of high-pressure frozen, freeze-substituted, resin-embedded, and thin-sectioned *Methylomirabilis* cells taken from a bioreactor enrichment culture. (A,B) Infected *Methylomirabilis* cells (white arrow in **A**) are in clusters among non-infected cells (black arrows in **A**). The bacteriophage (white arrow in **B**) has a hexagonal shape and an internal electron dense core. **(C,D)** The bacteriophages are organized in a highly packed formation (white arrow in **D**). The replication and assembly of bacteriophages causes the *Methylomirabilis* cell to swell and eventually the cell wall breaks (black arrows in **D**). **(E,F)** Lysed *Methylomirabilis* cell releasing the viral progeny. **(B,D,F)** are enlargements of **(A,C,E)**, respectively. Scale bars; 200 nm.

at both ends, indicating a full genome and a circular genomic arrangement. The 41 kb contig did not have overlapping end sequences, possibly representing a linear genome.

Phage Genome 1

In this genome, one head morphogenesis protein was found, involved in initial stages of head assembly (Hsiao and Black, 1978; Leiman et al., 2003). DNA replication enzymes are represented by one DNA polymerase and one DNA polymerase subunit. In addition, the two helicases may also take part in the replication process. The DNA packaging machinery in genome 1 includes three terminases. Compared to the other five annotated genomes, genome 1 was the only annotated genome that does not contain a clear hit to a tail protein. In addition, genome 1 encodes

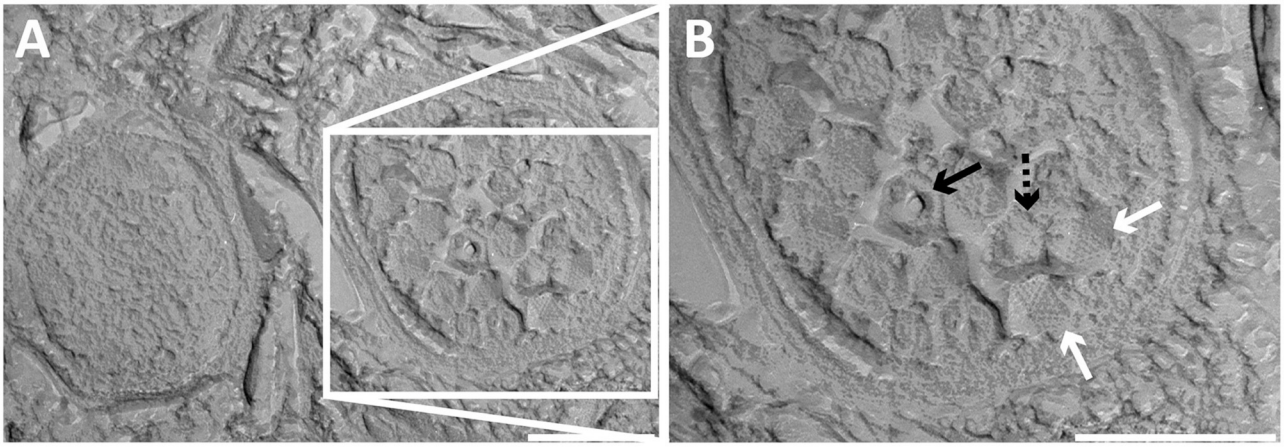


FIGURE 3 | Transmission electron micrographs of high-pressure frozen and freeze-etched *Methylomirabilis* cells taken from a bioreactor enrichment culture. (A) Cross-section of a non-infected (left) and infected (right) *Methylomirabilis* cell. **(B)** The bacteriophages (white arrows) contain a proteinaceous capsid. The capsid is made of triangular faces built by capsomeres. Concave (dashed black arrow) and convex (black arrow) printing of the internal core is visible in two of the viral particles. Scale bars; 200 nm.

for one putative ATPase, maybe involved in DNA packaging (Kondabagil et al., 2006), and two ORFs with a blast hit against a membrane protein. The genome encodes no less than 41 tRNAs. The acquisition of these tRNAs from the host might function in providing the phage with a similar GC content as the host, thereby facilitating the recruitment of the host DNA replication enzymes (Bailly-Bechet et al., 2007; Enav et al., 2012).

Phage Genome 2

Genome 2 encodes for one putative tail protein. No capsid-related genes could be annotated. The DNA replication machinery in genome 2 consists of one DNA polymerase, one helicase and one primase. Interestingly, a PKD (polycystic kidney disease) domain-containing protein was also annotated. These proteins may have a function in mediating phage-host contact (Fraser et al., 2006, 2007; Sathaliyawala et al., 2010) and BAM (bacteriophage adherence to mucus) mechanisms (Barr et al., 2013).

Phage Genome 3

Genome 3 contains one tail TMP (tape measure protein) and one head protein. Two helicases and one RNase (ribonuclease) H are the only three annotated genes that could be part of the DNA replication machinery. One chromosome partitioning protein similar to ParB was found in this genome. A ParB protein is also encoded by the *E. coli* phage P1, which can be present in the cell as a plasmid or integrated in the bacterial DNA (Abeles et al., 1985). The presence of a ParB-like protein in genome 3 may indicate that this bacteriophage might also be capable of integrating in the host genome or forming a plasmid. A putative phage-host interaction protein was also identified. This is a carbohydrate-binding protein with an Ig-like fold. Carbohydrate-binding domains have been found on tail spike proteins (Andres et al., 2010) and head spikes (Westbye et al., 2016), in both cases mediating the recognition of the target polysaccharides in the initial steps of infection.

Phage Genome 4

Genome 4 has three proteins involved in head assembly, being a prohead protease and two head morphogenesis proteins. Next to them in the genome two tail proteins are located (one tail TMP and one major tail subunit).

Phage Genome 5

Genome 5 encodes for a putative tail protein but no head proteins could be annotated. Most of the ORFs in this small genome are hypothetical proteins.

All annotated genomes show the putative ability to establish lysogeny. One transposase was found in genomes 2 and 5, two transposases in genomes 3 and 4 and three transposases in genome 1. These enzymes mediate unidirectional and site-specific recombination between two DNA sequences, one on the viral genome and one on the host genome (Groth and Calos, 2004). All genomes encode for one or two terminases (or translocase). These enzymes are responsible for packaging the genome inside the capsid (Catalano, 2000; Duffy and Feiss, 2002). HNH endonucleases found in double copy in genome 1 and in single copy in genomes 2 and 3 may also be involved in cutting the genome while DNA packaging is taking place (Moodley et al., 2012; Kala et al., 2014). Methyltransferases were also frequently found in single copy (genomes 1, 2, and 3) and double copy (genome 4). In bacteriophages methyltransferases are often thought to protect the viral genome from degradation by host-encoded edonucleases. However, few studies hypothesize a role in DNA replication and regulation both in lysogenic and lytic cycles (Murphy et al., 2013).

Genome Prediction of the *Methylomirabilis*-Infecting Bacteriophage

A main focus of the viral metagenomics in this study was to genomically identify the bacteriophage infecting *Methylomirabilis* cells (**Figures 2–4**). Based on the five viral genomes and over two thousand contig sequences obtained

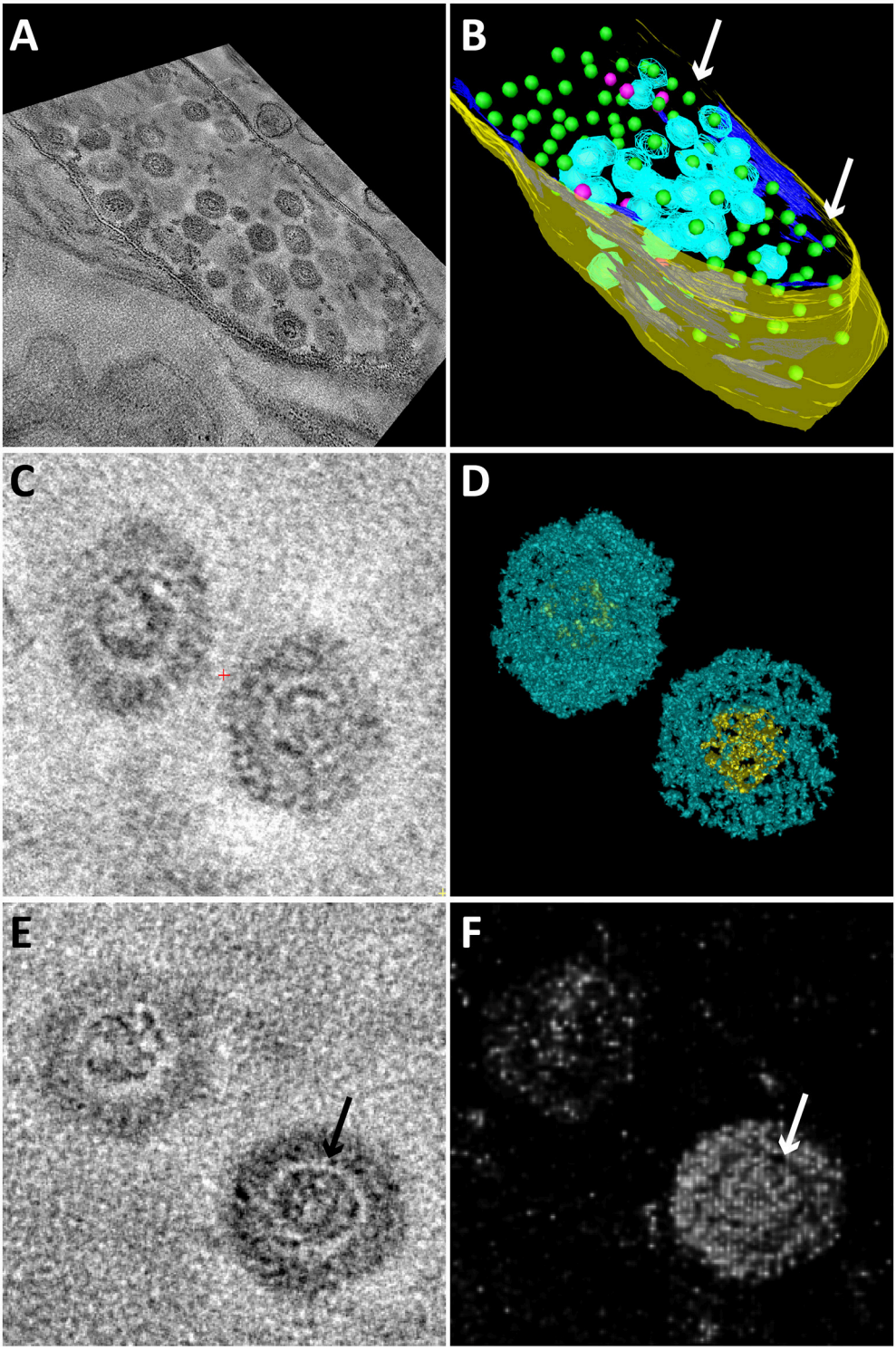


FIGURE 4 | Snapshots of electron tomograms and models of free and intracellular bacteriophages infecting *Methylomirabilis* cells. Tomogram **(A)** and model **(B)** of an infected *Methylomirabilis* cell. Most bacteriophages have the capsid (blue) assembled around the electron dense core (green). Some bacteriophages are still in the process of assembly and only consist of the electron dense core (pink). The cell is swollen and the cytoplasmic membrane (dark blue) is broken at many places (arrows). The cell wall (yellow) is still intact. All green electron dense cores were surrounded by a capsid, but not all capsids were modeled for reasons of clarity. Tomogram **(C)** and isosurface density model **(D)** of two free bacteriophages showing the icosahedral capsid (blue) and electron dense core (yellow). Tomogram **(E)** and Chimera model **(F)** showing two free bacteriophages. The electron dense core is enclosed by a putative membrane (arrows).

TABLE 2 | Genomic characteristics of the five putative complete viral genomes retrieved from the viral metagenome of the bacteriophage population in the *Methylomirabilis* bioreactor enrichment culture.

Genome	Genomic arrangement	Genome length (kb)	Number of reads	Depth	GC content (%)	# ORFs
1	circular	197	254,185	320.1	62.9	280
2	circular	86	213,396	266.8	54.1	103
3	circular	71	320,546	41.2	54.3	102
4	linear	41	672,323	86.4	57.2	54
5	circular	17	173,132	219.0	67.8	25

above, it is not directly clear which cellular hosts the recovered bacteriophages infect. Thus, to address this question we applied several computational approaches to predict this phage-host relationship from sequence signals (Edwards et al., 2015) that linked the assembled viral contigs to both the original genome and that of the new *Methylomirabilis* species. As described in Edwards et al. (2015), we calculated three main categories of phage-host prediction signals: genetic homology, similarity in CRISPR spacers, and similarity in oligonucleotide usage (Supplementary File 6). Genetic homology is expected to occur between the viral and bacterial genome sequences due to processes including transduction. This was identified at the nucleotide level (blastn, $E \leq 10^{-5}$) and at the protein level (tblastx, $E \leq 10^{-5}$), the former detecting only five short fragments, the latter detecting many more contigs including the long genomes 1–4. However, in all cases the matches between the bacteriophages and the *Methylomirabilis* genomes were based on relatively short, scattered regions of similarity with low sequence identity. As it was also showed by Edwards et al. (2015), the similarities in oligonucleotide usage profiles ($k = 2, 4, 6$) are not strong as phage-host predictors.

In the previously described *M. oxyfera* genome (Ettwig et al., 2010) one CRISPR array was present, which contained 23 spacers (Supplementary File 7). In the genome of a new *Methylomirabilis* species (Guerrero et al., in preparation) derived from the reactor from which the virome was sequenced, two CRISPR arrays were observed; one of which contained 26 spacers and the other one six spacers (Supplementary File 7). No significant nucleotide similarity was detected between these spacers and any of the 2094 possible viral contig sequences with fewer than 11/37 mismatches, which means these hits are indistinguishable from random noise (Edwards et al., 2015).

We also looked for signs of lysogeny in the *M. oxyfera* metagenome (Ettwig et al., 2010) and the new *Methylomirabilis* sp. metagenome (Guerrero et al., in preparation). The viral contigs were compared using BLASTn searching for contigs that were partly viral and partly bacterial. However, no such contigs were present in neither of the two metagenomes.

DISCUSSION

We analyzed the bacteriophage community in a bioreactor enrichment culture of the anaerobic methane oxidizing bacterium *Methylomirabilis* sp. Four different bacteriophage morphologies were observed inside the bioreactor system, one of which was observed to infect *Methylomirabilis* cells. From the viral metagenome data at least five putative complete viral genomes were assembled. Based on bioinformatics including CRISPR analyses, we could not identify the viral genome of the *Methylomirabilis* infecting bacteriophage.

Viral metagenome analysis of bioreactor enrichment cultures is challenging. In this case, the viral abundance was low and it required 13 L bioreactor material to obtain 0.2 ng viral DNA. We used a newly developed method for obtaining the (dsDNA) viral metagenome from the low amounts of viral DNA (Cremers et al., in preparation). In the end, 2094 contigs were assembled. Among these we identified five putative complete viral genomes. The annotation of these five genomes resulted in many hypothetical proteins. Metagenomics reads are often short and will not lead to improved annotation if viruses are not isolated and further experiments performed. Therefore, the viral sequences in the databases represent only a small fraction of the entire viral sequence space (Mokili et al., 2012; Dutilh, 2014). Nevertheless, we could in most cases identify putative genes involved in DNA processing and tail and capsid formation.

Electron microscopy analyses showed a bacteriophage infecting *Methylomirabilis* cells at a relatively low infection rate of 2.3%. The bacteriophage performed a lytic cycle as bacteriophages were replicated inside the host cell, causing the host to swell. In the end, the host cell lysed and released the viral particles. The bacteriophage had a putative icosahedral capsid and no tail. Inside the bacteriophage, a round electron dense core was visible, which most probably represented the highly packed genome. Electron tomography suggested the presence of a putative internal membrane, which appeared as an electron-light area surrounding the genome. Such an electron-light structure around a viral genome has not been observed for the other described morphotypes, therefore this could be an indication that this genome was enclosed by a structure, most likely a lipid membrane, as it was also shown for bacteriophages with similar morphology (Harrison et al., 1971; Mindich et al., 1982). Inside the infected *Methylomirabilis* cells incompletely assembled bacteriophages were observed constituted by only the electron dense core. The assembly of this bacteriophage might thus proceed by the following steps: membrane formation, packaging of the genome, and assembly of the capsid.

The presence of an internal membrane surrounding the viral genome has so far only been described for two bacteriophage families: *Tectiviridae* (four phages described and PRD1 as type species) and *Corticoviridae* (one phage described; PM2; King et al., 2012). These bacteriophages infect Gram-negative bacteria, have a relatively small dsDNA genome [10 kb circular,

FIGURE 5 | Schematic overview of the five putative complete viral genomes retrieved from the viral metagenome of the bacteriophage population in the *Methylomirabilis* bioreactor enrichment culture. Arrows with the same color represent genes encoding for the same putative functional category. Light gray arrows indicate hypothetical proteins. Black bars in genome 1 and genome 2 represent tRNAs. Appr-1-pp, ADP-ribose 1-phosphate phosphatase; BCMO, β-carotene 15, 15′-monooxygenase; HAD hydrolase, haloacid dehalogenase; MTase, methyltransferase; PKD-containing protein, polycystic kidney disease-containing protein; RNase H, ribonuclease H; tail TMP, tail tape measure protein.

Corticoviridae (Männistö et al., 1999); 15 kb linear, *Tectiviridae* (Bamford et al., 1991)] and have an icosahedral capsid [66 nm from facet to facet for *Tectiviridae* (Abrescia et al., 2004) and 55 nm from facet to facet for *Corticoviridae* (Kivelä et al., 2002)] with no tail. It was described that these bacteriophages obtain their internal membrane from the host cytoplasmic membrane using so called membrane proteins (Espejo and Canelo, 1968; Mindich et al., 1982). Also in the present study, the cytoplasmic membrane of infected *Methylomirabilis* cells was not intact anymore as was apparent from 3D electron tomography models. The similar ultrastructure of the *Methylomirabilis* bacteriophage to the *Corticoviridae* and *Tectiviridae* might indicate that it belongs to one of these families.

By using a bioinformatics approach, we tried to predict which viral contigs (2094) assembled from the viral metagenome data belonged to the *Methylomirabilis*- infecting bacteriophage. However, we were not able to obtain any convincing result. Homology searches yielded only short spurious hits, while the oligonucleotide usage profiles were too broad and did not provide a specific signal to any one of the longer contigs. In addition, the CRISPR arrays that were identified in the *Methylomirabilis* genomes did not have a clear hit to any of the assembled bacteriophage contigs. There are several possible explanations for the latter. First, since bacteriophage genomes with high similarity to one or more spacers in a CRISPR array are precluded from infection, CRISPR spacers are a focal point of positive selection in bacteriophage genome sequences, because the only bacteriophages that survive are the ones with mutations in the protospacer (Sun et al., 2013; Paez-Espino et al., 2015). Indeed, the closest match between CRISPR spacer and viral contigs already showed 11 mismatches. Second, the spacer composition of CRISPR arrays in a bacterial community can be very volatile due to ecological selection for bacterial strains with spacers acquired from the current bacteriophages in the environment (Rho et al., 2012; Koskella and Brockhurst, 2014). Thus, even though the *Methylomirabilis* genomes that were mined for CRISPR spacers were obtained from the same bioreactor (Ettwig et al., 2010; Guerrero et al., in preparation), not a single one of the CRISPR spacers was found in common. Taken together, computational phage-host signals did not result in a strong candidate for the bacteriophage infecting *Methylomirabilis* in this bioreactor.

We also investigated signs of lysogeny by in depth analyses of the *M. oxyfera* metagenome (Ettwig et al., 2010) and the new *Methylomirabilis* sp. metagenome (Guerrero et al., in preparation), looking for contigs that were partly bacterial and partly viral. However, no such sequences were detected. This could indicate that the bacteriophage does not perform a lysogenic cycle but only a lytic one. In the end, we can only speculate about which viral genome might belong to the bacteriophage observed to infect the *Methylomirabilis* cells. Unfortunately none of the five putative complete viral genomes matched all (indirect) criteria. The bacteriophage morphology (icosahedral capsid, no tail, genome surrounded by putative internal membrane) indicated a relation to the *Tectiviridae* and *Corticoviridae* families that have a relatively small genome size, although the bacteriophage could also belong to a yet-undiscovered family. The only small (complete) genome (genome 5–17 kb) contained a putative tail protein indicating that most likely it was not our tailless bacteriophage. There was one genome without an annotated tail gene but it was fairly large (genome 1–197 kb) which did not fit to the described genome size in these families. Other bacteriophages with an internal membrane and a genome size bigger than 15 kb have been described (Aalto et al., 2012; Rissanen et al., 2013) but they are not yet assigned to a viral family.

In our study we focused on DNA-containing viruses. One explanation why the applied bioinformatic approaches could not find any phage-host interaction signals could be because the *Methylomirabilis*-infecting bacteriophage might have a ssDNA or an RNA genome. However, ssDNA viruses and especially RNA viruses comprise only a small minority among prokaryotic viruses, dominated instead by dsDNA viruses (Koonin et al., 2015). Also, especially considering the low yield of viral DNA, we might not have retrieved the entire bacteriophage population from the bioreactor. Not until methods as phageFISH (Allers et al., 2013) are optimized to be used for non-model systems it remains difficult to link a phage to a host through non bioinformatic methods.

The *Methylomirabilis* infection rate was observed to be relatively low (2.3% of all infected *Methylomirabilis* cells) at different time points throughout the period of ∼1 year and did not affect bioreactor performance with respect to activity and growth. Indeed, the bacteriophage did not kill the *Methylomirabilis* population. Metagenomics analysis of the enrichment culture indicated the presence of a new *Methylomirabilis* species (Guerrero et al., in preparation), which was not present in the bioreactor when the genome of the original *M. oxyfera* species was published (Ettwig et al., 2010). An interesting speculation could be that the bacteriophage was specific for only the original *M. oxyfera* species (Ettwig et al., 2010) and in the end instigated the shift from the original to the new *Methylomirabilis* species. Indeed currently, the infection has almost disappeared from the enrichment culture, the published *M. oxyfera* 16S rRNA gene (NCBI, FP565575) cannot be detected anymore and the enrichment culture is dominated by the new *Methylomirabilis* species.

Bacteriophages can have profound effects on bacterial populations also in bioreactor enrichment systems. This has to be taken into account in the application of microorganisms in, for example, wastewater treatment systems or other industrial applications. In general, how bacteriophages can shape environmental or industrial microbial communities is still largely unexplored and is an exciting topic for further study.

AUTHOR CONTRIBUTIONS

LV and MJ designed the project. LG and LV designed the experiments. SG maintained and enriched the *Methylomirabilis* culture. LG and RM performed all TEM related experiments. LG performed the concentration of viral particles and DNA extraction. GC performed the library preparation and

sequencing. GC, BD, and HO designed, performed and analyzed the bioinformatics research. LG manually annotated the complete viral genomes. LV supervised the research. LG and LV performed data analysis, data interpretation and wrote the manuscript with input from GC, RM, SG, BD, MJ, and HO.

ACKNOWLEDGMENTS

We thank Jeremy J. Barr for critical reading of the manuscript and for the bacteriophage DNA isolation protocol; Christina Ferousi, Dave Allen, Felipe H. Coutinho, Diego D. Cambuy, and Bruno G. Andrade for practical assistance; Theo van Alen, and Daan Speth for advice on (all) and assistance with (TvA) running the ion torrent sequencer and analyzing the data; Boran Kartal for input for manuscript preparation; Geert-Jan Janssen and the General Instruments department for maintenance of TEM equipment. BD was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004. RM was supported by NWO Spinozapremie 2012 of MJ. MJ and LG are supported by ERC 339880, and MJ and GC also by OCW/NWO Gravitation grant (SIAM 024002002). HO is supported by ERC-AG 669371.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.01740/full#supplementary-material

REFERENCES

Aalto, A. P., Bitto, D., Ravantti, J. J., Bamford, D. H., Huiskonen, J. T., and Oksanen, H. M. (2012). Snapshot of virus evolution in hypersaline environments from the characterization of a membrane-containing *Salisaeta* icosahedral phage 1. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7079–7084. doi: 10.1073/pnas.1120174109

Abeles, A. L., Friedman, S. A., and Austin, S. J. (1985). Partitioning of unit-copy miniplasmids to daughter cells. III. The DNA sequence and functional organization of the P1 partition region. *J. Mol. Biol.* 185, 261–272. doi: 10.1016/0022-2836(85)90402-4

Abrescia, N. G. A., Cockburn, J. J. B., Grimes, J. M., Sutton, G. C., Diprose, J. M., Butcher, S. J., et al. (2004). Insight into assembly from structural analysis of bacteriophage PRD1. *Nature* 432, 68–74. doi: 10.1038/nature03056

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. doi: 10.1111/1462-2920.12100

Andres, D., Baxa, U., Hanke, C., Seckler, R., and Barbirz, S. (2010). Carbohydrate binding of Salmonella phage P22 tailspike protein and its role during host cell infection. *Biochem. Soc. Trans.* 38, 1386–1389. doi: 10.1042/BST0381386

Bailly-Bechet, M., Vergassola, M., and Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome Res.* 17, 1486–1495. doi: 10.1101/gr.6649807

Bamford, J. K. H., Hänninen, A. L., Pakula, T. M., Ojala, P. M., Kalkkinen, N., Frilander, M., et al. (1991). Genome organization of membrane-containing bacteriophage PRD1. *Virology* 183, 658–676. doi: 10.1016/0042-6822(91)90995-N

Barr, J. J., Auro, R., Furlan, M., Whiteson, K. L., Erb, M. L., Pogliano, J., Stotland, A., Wolkowicz, R., et al. (2013). Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 10771–10776. doi: 10.1073/pnas.1305923110

Barr, J. J., Slater, F. R., Fukushima, T., and Bond, P. L. (2010). Evidence for bacteriophage activity causing community and performance changes in a phosphorus-removal activated sludge. *FEMS Microbiol. Ecol.* 74, 631–642. doi: 10.1111/j.1574-6941.2010.00967.x

Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C., and Brown, C. M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* 10, 817–827. doi: 10.4161/rna.24046

Borrel, G., Colombet, J., Robin, A., Lehours, A. C., Prangishvili, D., and Sime-Ngando, T. (2012). Unexpected and noved putative viruses in the sediments of a deep-dark permanently anoxic freshwater habitat. *ISME J.* 6, 2119–2127. doi: 10.1038/ismej.2012.49

Breitbart, M., Salamon, P., Andersen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Breitbart, M., Thompson, L. R., Suttle, C. A., and Sullivan, B. (2007). Exploring the vast diversity of marine viruses. *Oceanography* 20, 135–139. doi: 10.5670/oceanog.2007.58

Brum, J. R., Schenck, R. O., and Sullivan, M. B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* 7, 1738–1751. doi: 10.1038/ismej.2013.67

Camacho, C., Coulouris, G., Avagyan, V., MA, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421–430. doi: 10.1186/1471-2105-10-421

Catalano, C. E. (2000). The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. *Cell. Mol. Life Sci.* 57, 128–148. doi: 10.1007/s000180050503

Colombet, J., and Sime-Ngando, T. (2012). "Use of PEG, Polyethylene glycol, to characterize the diversity of environmental viruses," in *Current Microscopy Contributions to Advances in Science and Technology*, ed A. Méndez-Vilas (Badajoz: Formatex Research Center), 316–322.

Cunningham, B. R., Brum, J. R., Schwenck, S. M., Sullivan, M. B., and John, S. G. (2015). An inexpensive, accurate and precise wet-mount method for enumerating aquatic viruses. *Appl. Environ. Microbiol.* 81, 2995–3000. doi: 10.1128/AEM.03642-14

De Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., et al. (2006). ScanProsite: detection of PROSITE signature matches and Pro-Rule-associated functional and structural residues in proteins. *Nucleic Acids Res.* 34, 362–365. doi: 10.1093/nar/gkl124

Deutzmann, J. S., Stief, P., Brandes, J., and Schink, B. (2014). Anaerobic methane oxidation coupled to denitrification is the dominant methane sink in a deep lake. *Proc. Natl. Acad. Sci. U.S.A.* 111, 18273–18278. doi: 10.1073/pnas.1411617111

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of the microbial genome sequence signatures. *Genome Biol.* 10:R85. doi: 10.1186/gb-2009-10-8-r85

Dimmock, N. I., Easton, A. J., and Leppard, K. N. (2007). *Introduction to Modern Virology*. Coventry: Department of Biological Sciences, University of Warwick.

Duffy, C., and Feiss, M. (2002). The large subunit of bacteriophage λ's terminase plays a role in DNA translocation and packaging termination. *J. Mol. Biol.* 316, 547–561. doi: 10.1006/jmbi.2001.5368

Dutilh, B. E. (2014). Metagenomic ventures into outer sequence space. *Bacteriophage* 4:e979664. doi: 10.4161/21597081.2014.979664

Edwards, R. A., McNair, K., Faust, K., Raes, J., and Dutilh, B. E. (2015). Computational approaches to predict phage-host relationships. *FEMS Microbial. Rev.* 40, 258–272. doi: 10.1093/femsre/fuv048

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

Enav, H., Béjà O., and Mandel-Gutfreund, Y. (2012). Cyanophage tRNAs may have a role in cross-infectivity of oceanic *Prochlorococcus* and *Synechococcus* hosts. *ISME J.* 6, 619–628. doi: 10.1038/ismej.2011.146

Espejo, R. T., and Canelo, E. S. (1968). Origin of phospholipid in bacteriophage PM2. *J. Virol.* 2, 1235–1240.

Ettwig, K. F., Butler, M. K., Le Paslier, D., Pelletier, E., Mangenot, S., Kuypers, M. M. M., et al. (2010). Nitrate-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* 464, 543–550. doi: 10.1038/nature08883

Ettwig, K. F., van Alen, T., van de Pas-Schoonen, K. T., Jetten, M. S. M., and Strous, M. (2009). Enrichment and molecular detection of denitrifying methanotrophic bacteria of the NC10 phylum. *Appl. Environ. Microbiol.* 75, 3656–3662. doi: 10.1128/AEM.00067-09

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science* 320, 1034–1039. doi: 10.1126/science.1153213

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, 279–285. doi: 10.1093/nar/gkv1344

Fraser, J. S., Maxwell, K. L., and Davidson, A. R. (2007). Immunoglobulin-like domains on bacteriophage: weapons of modest damage? *Curr. Opin. Microbiol.* 10, 382–387. doi: 10.1016/j.mib.2007.05.018

Fraser, J. S., Yu, Z., Maxwell, K. L., and Davidson, A. R. (2006). Ig-like domains on bacteriophages: a tale of promiscuity and deceit. *J. Mol. Biol.* 359, 496–507. doi: 10.1016/j.jmb.2006.03.043

Groth, A. C., and Calos, M. P. (2004). Phage integrases: biology and applications. *J. Mol. Biol.* 335, 667–678 doi: 10.1016/j.jmb.2003.09.082

Guo, P., El-Gohary, Y., Prasadan, K., Shiota, C., Xiao, X., Wiersch, J., et al. (2012). Rapid and simplified purification of recombinant adeno-associated virus. *J. Virol. Methods* 183, 139–146. doi: 10.1016/j.jviromet.2012.04.004

Harrison, S. C., Caspar, D. L. D., Camerini-Otero, R. D., and Franklin, R. M. (1971). Lipid and protein arrangement in bacteriophage PM2. *Nature New Biol.* 229, 197–201. doi: 10.1038/newbio229197a0

Hatfull, G. F. (2008). Bacteriophage genomics. *Curr. Opin. Microbiol.* 11, 447–453. doi: 10.1016/j.mib.2008.09.004

Hsiao, C. L., and Black, L. W. (1978). Head morphogenesis of bacteriophage T4. II. The role of gene 40 in initiating prehead assembly. *Virology* 91, 15–25. doi: 10.1016/0042-6822(78)90351-3

Hu, B. L., Shen, L. D., Lian, X., Zhu, Q., Liu, S., Huang, Q., et al. (2014). Evidence for nitrite-dependent anaerobic methane oxidation as a previously overlooked microbial methane sink in wetlands. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4495–4500. doi: 10.1073/pnas.1318393111

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. doi: 10.1371/journal.pone.0057355

John, S. G., Mendez, C. B., Deng, L., Poulos, B., Kauffman, A. K. M., Kern, S., et al. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* 3, 195–202. doi: 10.1111/j.1758-2229.2010.00208.x

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. doi: 10.1093/nar/gkn201

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kala, S., Cumby, N., Sadowski, P. D., Zafar Hyder, B., Kanelis, V., Davidson, A. R., et al. (2014). HNH proteins are a widespread component of phage DNA packaging machines. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6022–6027. doi: 10.1073/pnas.1320952111

King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (2012). *Virus Taxonomy. Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses*. San Diego, CA: Elsevier Academic Press.

Kivelä, H. M., Kalkkinen, N., and Bamford, D. H. (2002). Bacteriophage PM2 has a protein capsid surrounding a spherical proteinaceous lipid core. *J. Virol.* 76, 8169–8178. doi: 10.1128/JVI.76.16.8169-8178.2002

Kondabagil, K. R., Zhang, Z., and Rao, V. B. (2006). The DNA translocating ATPase of bacteriophage T4 packaging motor. *J. Mol. Biol.* 363, 786–799. doi: 10.1016/j.jmb.2006.08.054

Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 479–480, 2–25. doi: 10.1016/j.virol.2015.02.039

Koskella, B., and Brockhurst, M. A. (2014). Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiol. Rev.* 38, 916–931. doi: 10.1111/1574-6976.12072

Koskella, B., and Parr, N. (2015). The evolution of bacterial resistance against bacteriophages in the horse chestnut phyllosphere is general across both space and time. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370:20140297. doi: 10.1098/rstb.2014.0297

Kremer, J. R., Mastronarde, D. N., and McIntosh, J. R. (1996). Computer visualization of three-dimentional image data using IMOD. *J. Struct. Biol.* 116, 71–76. doi: 10.1006/jsbi.1996.0013

Kunin, V., He, S., Warnecke, F., Peterson, S. B., Garcia Martin, H., Haynes, M., et al. (2008). A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome Res.* 18, 293–297. doi: 10.1101/gr.6835308

Leiman, P. G., Kanamaru, S., Mesyanzhinov, V. V., Arisaka, F., and Rossmann, M. G. (2003). Structure and morphogenesis of bacteriophage T4. *Cell. Mol. Life Sci.* 60, 2356–2370. doi: 10.1007/s00018-003-3072-1

Letunic, I., Doerks, T., and Bork, P. (2014). SMART: recent updates, new developments and status 2015. *Nucleic Acids Res.* 43, 257–260. doi: 10.1093/nar/gku949

Luesken, F. A., van Alen, T. A., van der Biezen, E., Frijters, C., Toonen, G., Kampman, C., et al. (2011). Diversity and enrichment of nitrite-dependent anaerobic methane oxidizing bacteria from wastewater sludge. *Environ. Biotechnol.* 92, 845–854. doi: 10.1007/s00253-011-3361-9

Männistö, R. H., Kivelä, H. M., Paulin, L., Bamford, D. H., and Bamford, J. K. H. (1999). The complete genome sequence of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* 262, 355–363. doi: 10.1006/viro.1999.9837

Mastronarde, D. N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* 152, 36–51. doi: 10.1016/j.jsb.2005.07.007

Middelboe, M., Jacquet, S., and Weinbauer, M. (2008). Viruses in freshwater ecosystems: an introduction to the exploration of viruses in new aquatic habitats. *Freshwater Biol.* 53, 1069–1075. doi: 10.1111/j.1365-2427.2008.02014.x

Mindich, L., Bamford, D., McGraw, T., and Mackenzie, G. (1982). Assembly of bacteriophage PRD1: particle formation with wild-type and mutant viruses. *J. Virol.* 44, 1021–1030.

Mokili, J. L., Rohwel, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Mollenhauer, H. H. (1964). Plastic embedding mixtures for use in electron microscopy. *Stain Technol.* 39, 111–114.

Moodley, S., Maxwell, K. L., and Kanelis, V. (2012). The protein gp74 from the bacteriophage HK97 functions as a HNH endonuclease. *Protein Sci.* 21, 809–818. doi: 10.1002/pro.2064

Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., and van Sinderen, D. (2013). Bacteriophage orphan DNA methyltransferases: insight from their bacterial origin, function and occurrence. *Appl. Environ. Microbiol.* 79, 7547–7555. doi: 10.1128/AEM.02229-13

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., et al. (2014). *IPCC, 2014: Climate Change 2014: Synthesis Report*. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

Paez-Espino, D., Eloe-Fadrosh, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., et al. (2016). Uncovering Earth's virome. *Nature* 536, 425–430. doi: 10.1038/nature19094

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B. C., Barrangou, R., et al. (2015). CRISPR immunity drives rapid phage genome evolution in *Streptococcus thermophilus*. *mBio* 6:e00262-15. doi: 10.1128/mBio.00262-15

Pagni, M., Ioannidis, V., Cerutti, L., Zahn-Zabal, M., Jongeneel, C. V., Hau, J., et al. (2007). MyHits: improvements to an interactive resource for analyzing protein sequences. *Nucleic Acids Res.* 35, 433–437. doi: 10.1093/nar/gkm352

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSFChimera—a visualization system for

exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084

Rho, M., Wu, Y. W., Tang, H., Doak, T. G., and Ye, Y. (2012). Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.* 8:e1002441. doi: 10.1371/journal.pgen.1002441

Rissanen, I., Grimes, J. M., Pawlowski, A., Mäntynen, S., Harlos, K., Bamford, J. K. H., et al. (2013). Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage. *Structure* 21, 718–726. doi: 10.1016/j.str.2013.02.026

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., et al. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* 16, 944–945. doi: 10.1093/bioinformatics/16.10.944

Sathaliyawala, T., Islam, M. Z., Li, Q., Fokine, A., Rossmann, M. G., and Rao, V. B. (2010). Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. *Mol. Microbiol.* 77, 444–455. doi: 10.1111/j.1365-2958.2010.07219.x

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153

Shapiro, O. H., Kushmaro, A., and Brenner, A. (2010). Bacteriophage predation regulates microbial abundance and diversity in a full-scale bioreactor treating industrial wastewater. *ISME J.* 4, 327–336. doi: 10.1038/ismej.2009.118

Shi, Y., Hu, S., Lou, J., Lu, P., Keller, J., and Yuan, Z. (2013). Nitrogen removal from wasterater by coupling anammox and methane-dependent denitrification in a membrane biofilm reactor. *Environ. Sci. Technol.* 47, 11577–11583. doi: 10.1021/es402775z

Skennerton, C. T., Imelfort, M., and Tyson, G. W. (2013). Crass: identification and reconstruction of CRISPS from unassembled metagenomic data. *Nucleic Acids Res.* 41:e105. doi: 10.1093/nar/gkt183

Sulcius, S., Staniulis, J., and Paškauskas, R. (2011). Morphology and distribution of phage-like particles in a eutrophic boreal lagoon. *Oceanologia* 53, 587–603. doi: 10.5697/oc.53-2.587

Sun, C. L., Barrangou, R., Thomas, B. C., Horvath P., Fremaux, C., and Banfield, J. F. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol.* 15, 463–470. doi: 10.1111/j.1462-2920.2012.02879.x

Suttle, C. A. (2005). Viruses in the sea. *Nature* 437, 356–361. doi: 10.1038/nature04160

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

Szpara, M. L., Tafuri, Y. R., and Enquist, L. W. (2011). Preparation of viral DNA from nucleocapsids. *J. Vis. Exp.* 54, 1–6. doi: 10.3791/3151

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

Ultsch, A., and Mörchen, F. (2005). *ESOM-Maps: Tool for Clustering, Visualization, and Classification with Emergent SOM*. Technical Report Department of Mathematics and Computer Science, University of Marburg, Germany.

Walther, P., and Ziegler, A. (2002). Freeze substitution of high-pressure frozen samples: the visibility of biological membranes is improved when the substitution medium contains water. *J. Microsc.* 208, 3–10. doi: 10.1046/j.1365-2818.2002.01064.x

Weinbauer, M. G., and Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* 6, 1–11. doi: 10.1046/j.1462-2920.2003.00539.x

Westbye, A. B., Kuchinski, K., Yip, C. K., and Beatty, J. T. (2016). The gene transfer agent RcGTA contains head spikes needed for binding to the *Rhodobacter capsulatus* polysaccharide cell capsule. *J. Mol. Biol.* 428, 477–491. doi: 10.1016/j.jmb.2015.12.010

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64, 69–114. doi: 10.1128/MMBR.64.1.69-114.2000

Wu, C. H., Yeh, L. S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., et al. (2003). The protein information resource. *Nucleic Acids Res.* 31, 345–347. doi: 10.1093/nar/gkg040

Wu, M. L., van Teeseling, M. C. F., Willems, M. J. R., van Donselaar, E. G., Klingl, A., Rachel, R., et al. (2012). Ultrastructure of the denitrifying methanotroph '*Candidatus Methylomirabilis oxyfera*', a novel polygon-shaped bacterium. *J. Bacteriol.* 194, 284–291. doi: 10.1128/JB.05816-11

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

frontiers in Microbiology

METHODS
published: 08 May 2015
doi: 10.3389/fmicb.2015.00381

Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes

Ramy K. Aziz[1,2,3]*, *Bhakti Dwivedi*[4†], *Sajia Akhter*[1†], *Mya Breitbart*[4] and *Robert A. Edwards*[1,3]*

[1] Department of Computer Science, San Diego State University, San Diego, CA, USA, [2] Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt, [3] Computing, Environment, and Life Sciences, Argonne National Laboratory, Argonne, IL, USA, [4] College of Marine Science, University of South Florida St. Petersburg, St. Petersburg, FL, USA

Phages are the most abundant biological entities on Earth and play major ecological roles, yet the current sequenced phage genomes do not adequately represent their diversity, and little is known about the abundance and distribution of these sequenced genomes in nature. Although the study of phage ecology has benefited tremendously from the emergence of metagenomic sequencing, a systematic survey of phage genes and genomes in various ecosystems is still lacking, and fundamental questions about phage biology, lifestyle, and ecology remain unanswered. To address these questions and improve comparative analysis of phages in different metagenomes, we screened a core set of publicly available metagenomic samples for sequences related to completely sequenced phages using the web tool, Phage Eco-Locator. We then adopted and deployed an array of mathematical and statistical metrics for a multidimensional estimation of the abundance and distribution of phage genes and genomes in various ecosystems. Experiments using those metrics individually showed their usefulness in emphasizing the pervasive, yet uneven, distribution of known phage sequences in environmental metagenomes. Using these metrics in combination allowed us to resolve phage genomes into clusters that correlated with their genotypes and taxonomic classes as well as their ecological properties. We propose adding this set of metrics to current metaviromic analysis pipelines, where they can provide insight regarding phage mosaicism, habitat specificity, and evolution.

Keywords: virus, bacteriophage, genomics, metagenomics, ecology

OPEN ACCESS

Edited by:
Alejandro Reyes, Universidad de los Andes, Colombia

Reviewed by:
Guilherme Corrêa De Oliveira, Fundação Oswaldo Cruz - Fiocruz-Minas, Brazil
Martha Josefina Vives, Universidad de los Andes, Colombia
Alejandro Caro-Quintero, Corpoica, Colombia

***Correspondence:**
Ramy K. Aziz, Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo 11562, Egypt
ramy.aziz@gmail.com;
Robert A. Edwards, Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182, USA
redwards@mail.sdsu.edu

† ***Present Address:***
Bhakti Dwivedi, The Winship Cancer Institute of Emory University, Atlanta, USA
Sajia Akhter, Department of bioengineering, Stanford University, Stanford, USA

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *12 January 2015*
Accepted: *13 April 2015*
Published: *08 May 2015*

Citation:
Aziz RK, Dwivedi B, Akhter S, Breitbart M and Edwards RA (2015) Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. Front. Microbiol. 6:381. doi: 10.3389/fmicb.2015.00381

Introduction

Viruses are the most abundant and diverse nucleic acid-based entities on Earth (Weinbauer, 2004; Edwards and Rohwer, 2005; Thurber, 2009). Their population densities are estimated to be 10^9 per gram of soil (Williamson et al., 2005), 10^7 per ml of seawater (Bergh et al., 1989; Wommack and Colwell, 2000), and 10^{31} planet-wide (Whitman et al., 1998). There are approximately 10 times as many viruses as the combined number of all cellular organisms, and most viruses are bacteriophages (phages), viruses that infect bacteria (Edwards and Rohwer, 2005).

Although phages play critical biological and ecological roles (Weinbauer, 2004; Abedon, 2009; Breitbart, 2012) and are the cornerstone of major molecular biology discoveries, the current number of completely sequenced phage genomes lags behind those of cellular organisms, and information about the abundance and distribution of these sequenced phage genomes in various ecosystems remains limited. A striking example of how little we know about phage abundance and distribution is that two prevalent phages with near-universal distribution in the oceans (Zhao et al., 2013) and human feces (Dutilh et al., 2014) were part of the unknown biological dark matter until only recently.

Traditional experimental strategies tend to underestimate phage diversity, mostly because culture-based methods miss the majority of phages. Furthermore, the actual abundance of phage nucleic acids in the environment is greater than that calculated from phage particle enumeration, since phage nucleic acids can be either packaged in free phage particles, or concealed as prophages within bacterial and archaeal genomes (Edwards and Rohwer, 2005; Angly et al., 2006). On the other hand, sequence-based strategies, notably the metagenomics technologies developed in the past decade (Breitbart et al., 2002; Breitbart and Rohwer, 2005), have revolutionized phage ecology (e.g., Breitbart et al., 2003; Angly et al., 2006; Thurber et al., 2009; Belcaid et al., 2010; Rodriguez-Brito et al., 2010; Swanson et al., 2011; Mizuno et al., 2013; Martinez Martinez et al., 2014). Despite those major advances, systematic surveys of phage genes and genomes in available metagenomes remain scarce partly because of the lack of well-established mathematical methods or metrics that define various aspects of phage distribution, abundance, and gene coverage.

Here we set out to define and deploy a set of metrics to better describe multiple dimensions of phage ecological properties. To this end, we implemented a scaffolding approach through the Phage Eco-Locator web-tool [URL: http://www.phantome.org/eco-locator (Aziz et al., 2011)], combined with a multidimensional set of metrics to enable a systematic analysis of phages in nature. To demonstrate these metrics and explore their significance, relevance, and applicability, this manuscript describes the abundance, ubiquity, diversity, and habitat-specificity of 588 completely sequenced viruses in 296 metagenomes from various ecosystems (**Figure S1**). The metrics described here can be used, individually or in combination, for the analysis of any set of metagenomes vs. any set of phages, regardless of the analysis platform, as long as the number of phage hits per metagenomic sample is available.

Methods

Input Sequence Data (Figure S1)

(1) **Viral genomic data**. Viral genome sequences (582 phages, four of which contain three-segment genomes, i.e., three contigs each, as well as six archaeal viruses) were directly downloaded from the PhAnToMe database (URL: http://www.phantome.org/Downloads).

(2) **Metagenomic data**. The 296 metagenomic data sets used for testing the methods consist of unassembled metagenomic sequences that had been originally annotated or re-annotated in the Metagenomics RAST server–version 3 (Meyer et al., 2008), then were cleaned up (Schmieder et al., 2010; Schmieder and Edwards, 2011a) or dereplicated (Schmieder and Edwards, 2011b) and deposited in MyMgDB (URL: http://edwards.sdsu.edu/cgi-bin/mymgdb/show.cgi). The sources of these metagenomic data sets and other metadata used in the analysis are provided in supporting online material (**Table S1**). Bacterial community structure in the same metagenomic data sets was analyzed by FOCUS (Silva et al., 2014).

Phage Eco-Locator

Phage Eco-Locator (URL: http://www.phantome.org/eco-locator) is a Web interface, written in a combination of PERL, GnuPlot, and CGI scripts, that stores and visualizes precomputed tBLASTX (Altschul et al., 1997) results using dereplicated metagenomic DNA sequence reads as BLAST queries against a database of complete phage genomes (Aziz et al., 2011). For this study, a tBLASTX match to a phage sequence was considered significant if it had an *E*-value $\leq 10^{-5}$. The web tool allows examining matches with *E*-value threshold of 0.01 as well.

Metrics Describing Phage Abundance and Distribution in Ecosystems

As indicated in the Introduction section, this work was launched with the goal of defining and testing metrics that describe different aspects of phage ecological properties, through the interpretation of phage metagenomic recruitment plots, to compare the abundance and distribution of sequences from different phages in various metagenomes, and also compare different metagenomic samples based on their phage content and abundance.

Those metrics fall into two major groups:

(i) Metagenome-level metrics: Metrics comparing different metagenomic data sets based on phage abundance and distribution (**Table 1**).

(ii) Phage genome-level metrics:

  (a) Metrics that describe a specific phage's abundance and distribution (on the genome level) (**Table 2**, **Figure 1**).

  (b) Metrics that describe the pattern of abundance, distribution, and coverage of different genes or segments within a specific phage genome in metagenomic data sets (**Table 3**, **Figure 2**).

(i) Metagenome-level metrics (Table 1). The following metrics are defined to provide a comparison between different metagenomes based on the abundance and distribution of sequences similar to characterized phages that they contain.

First, *all* metagenomic sequence reads with significant tBLASTX hits to phage sequences were collected from Eco-Locator recruitment plots and stored for further calculations. Those values were counted and defined as nHits. Default significance thresholds were set at BLAST *E*-values of 10^{-5}.

TABLE 1 | Metrics used to describe and compare different metagenomes based on their phage content (metagenome-level metrics).

Parameter	Definition/Calculation	Range	Significance/Interpretation/Limitations
IN A GIVEN METAGENOME Y			
Abundance index (AI) of phage X	nHits of phage X/size of metagenome Y (Mbp)	0–1.244	This value describes the fraction of a metagenome library that matches a given phage genome. Dividing the number of sequence hits by the metagenome size (in millions of basepairs) permits comparison of different metagenomic samples.
Total AI	Σ nHits of a set of phages/size of metagenome Y (Mbp)	4.067–28.859	This value reflects the abundance of all sequences with similarity to phages in a metagenomic library. *Limitations*: sensitive to outlier AI values (contaminants, sequencing artifacts, unusually large number of hits), i.e., false positive hits of a single phage can artificially inflate this value.
Median AI (AI_{50})	AI of the 50th percentile phage genome	0–3.061	This value gives an indication of the abundance of sequences with similarity to phages within a metagenomic library and is less sensitive to outliers than Total AI; however, it may underestimate real differences between samples (e.g., if more than half of the phage genomes have no sequence similarities to a metagenomic library, AI_{50} will be zero regardless of whether the total abundance of the remaining phage genomes is high or low).
nPhages (richness)	Number of phage genomes which match at least one sequence read in metagenome Y	8–487	This value is a proxy for *richness* of phage types within the metagenomic sample. While this value may overestimate the number of phage types within the tested sample, it can be used to compare sequence diversity between the tested metagenomic samples.
Shannon Diversity Index	$H = -\Sigma\, p_i \ln p_i$ where p_i is the proportion of sequence hits to the i^{th} phage genome relative to all phage genome hits within the metagenome	2.061–5.813	This value (Shannon, 1948) is an indication of the *diversity* of phage sequences within a metagenomic sample, but is not an accurate estimation of phage species diversity [which is beyond the focus of this paper and is to be calculated by other tools, e.g., PHACCS (Angly et al., 2005) or Shotgun UNIFRAC (Caporaso et al., 2011)].
Shannon E (evenness)	E = H/ln nPhages	0.008–0.258	This value describes the *evenness* of distribution of phage genomes. When Shannon *E*-value = 1, all genomes are equally represented; a Shannon *E*-value that is closer to zero reflects that an uneven distribution where some genomes are much more represented than others.

Next, an *abundance index* (AI) was calculated for each metagenome. For a given metagenome, the AI was defined as the number of hits to phage genomes (nHits) normalized to the metagenome size in millions of base pairs.

$$\text{AI} = \text{nHits/metagenome size, Mbp}$$

Subsequently, a *total abundance index* was defined for each metagenome to express the overall abundance of sequences with similarities to characterized phage genomes in that metagenome.

$$\text{Total abundance index (of all phage genomes) per metagenome} = \Sigma(\text{nHits/metagenome size, Mbp})$$

Because of the high variability of phage types in different ecosystems, the total AI defined above is highly sensitive to outliers, and thus the *median AI* of sequences with similarities to characterized phage genomes per metagenome was calculated as another useful value to compare metagenomes and reflect their phage content.

In addition to AI and median AI, which reflect phage-like metagenomic fragment counts, we also used some commonly used ecological biodiversity parameters such as richness, diversity, and evenness, described elsewhere (Shannon, 1948 disambiguated in Spellerberg and Fedor, 2003).

A full list of metagenome-level metrics, and the significance of each, is provided in **Table 1**.

(ii) Phage genome-level metrics.

(a) **Inter-phage properties (Table 2).** For comparison of phage genomes, a *phage abundance index* (PAI) was defined for each phage and calculated as the number of metagenomic sequence fragments assignable to that phage genome normalized to the genome size

$$\text{PAI} = \Sigma\ \text{AI/Phage genome length (Kbp)}$$

Because PAI depends on summing up all available metagenomic sequences that are similar to a particular phage, this value reflects

TABLE 2 | Metrics used to describe phage ecological features at the genome level.

Parameter	Definition/Calculation	Range	Significance/Interpretation/Limitations
PHAGE DISTRIBUTION METRICS (GENOME-LEVEL METRICS): FOR A GIVEN PHAGE X			
Phage abundance index (PAI)	Σ AI of phage X (hits per Mbp)/length of phage X (in Kbp)	0–194.84	This value describes the abundance of a phage in a set of environments. Normalizing the AI of each phage genome to the genome length allows the comparison of different phages. This normalization is useful for most phages; however, it might artificially inflate PAI value if the phage genome is significantly smaller than the median genome size, which is ∼41 Kbp (e.g., microviruses, with 4 Kbp genomes)
nMG	Number of metagenomes with hits to phage X	0–293	This value reflects the ubiquity of a particular phage genome. A high nMG suggests that a phage genome (or part of it) is universally distributed or cosmopolitan; a low nMG suggests that the phage is localized or ecologically limited (i.e., specific to one or a few habitats).
PAI_{50}	Median AI of phage X in all tested metagenomes/length of phage X	0–0.13	This value is another indication of the abundance of a phage genome in different metagenomic samples and is less sensitive to outliers. It is also dependent on the ubiquity of a phage genome since PAI_{50} of phage genomes present in fewer than half samples, for example, will be zero, even if these genomes have a high PAI.
Abund. CV (Coefficient of variation)	StDev/Mean AI of phage X	0.86–17.20	This value reflects the spread or variation of AIs of a given phage among metagenomes. A large CV suggests that a phage genome has extreme AIs while a small CV suggests uniform AI values (but doesn't give information on their magnitude).

*Representative examples of each value are given in **Figure 1**.*

a phage's overall abundance in a set of ecosystems, but provides little information about the pattern of its distribution, since a very high PAI may be contributed by an overabundance in a small number of metagenomes (nMGs).

Instead, an estimation of the distribution of a certain phage in a given set of ecosystems may be expressed as a simple count of the nMGs with significant BLAST hits (E-value $< 10^{-5}$) to a given phage genome. With a large nMGs from distinct ecosystems, nMG can be reliably used as a proxy for phage ubiquity in nature. In addition to counting metagenomes with hits to a given phage, we calculate the median PAI (PAI_{50}), an estimator of both the abundance and ubiquity of that phage in nature (**Table 2**).

Combining PAI, PAI_{50}, and nMG in comparisons between different phages provides a good multidimensional picture of phage distribution in nature, balancing abundance and ubiquity, as those two values do not necessarily correlate (**Figure 3A**). Those values, however, do not tell much about the uniformity of a phage's distribution among ecosystems. A phage with high PAI and low nMG is expected to have a highly variable distribution pattern in nature. This variability can be expressed as the *abundance coefficient of variation* (Abundance CV), representing the *data spread* of a phage genome's AI across metagenomic data sets, where CV is the standard deviation divided by the mean.

$$\text{Abundance CV} = \sigma \text{ AI}/\text{mean AI}$$

(b) **Intra-phage properties (Table 3).** Fragment recruitment plots and genome coverage maps are quite popular in analyzing metagenomic data; yet, a wealth of information encoded within those plots remains unexplored. Phage Eco-Locator, like other common metavirome analysis tools, e.g., MG-RAST (Meyer et al., 2008) and MetaVir (Roux et al., 2011), displays fragment recruitment plots, in which each metagenomic fragment is aligned to the corresponding genomic segment, as well as genome coverage density plots, in which each metagenomic sequence is cumulatively plotted against a phage genome scaffold, at a nucleotide resolution.

Coverage density plots provide a quick visual estimate of phage sequence conservation and distribution in a given metagenome. However, these plots are often biased by the presence of short sequences that are highly abundant (e.g., short repeats indicative of transposons or insertion sequences). Several mathematical values are suggested here to estimate different features of sequence coverage along a phage scaffold (**Table 3**). For example, *coverage density* may be measured as the area under the curve (AUC) normalized to the genome length (in nucleotides). For a certain phage, the total (or cumulative) coverage density in a large set of metagenomes may be further normalized to the nMGs with hits to that phage. As with other metrics, coverage density or cumulative coverage density is sensitive to outliers. Thus, *median coverage density* can be used to reflect the homogeneity of phage genome coverage in metagenomic samples.

In addition to coverage density, recruitment can be described by the uniformity, regularity, or continuity of sequence coverage over the entire genome length. Uniformity may be measured in various ways. One way is to simply estimate the percentage of a phage genome that recruits metagenomic reads (with possible

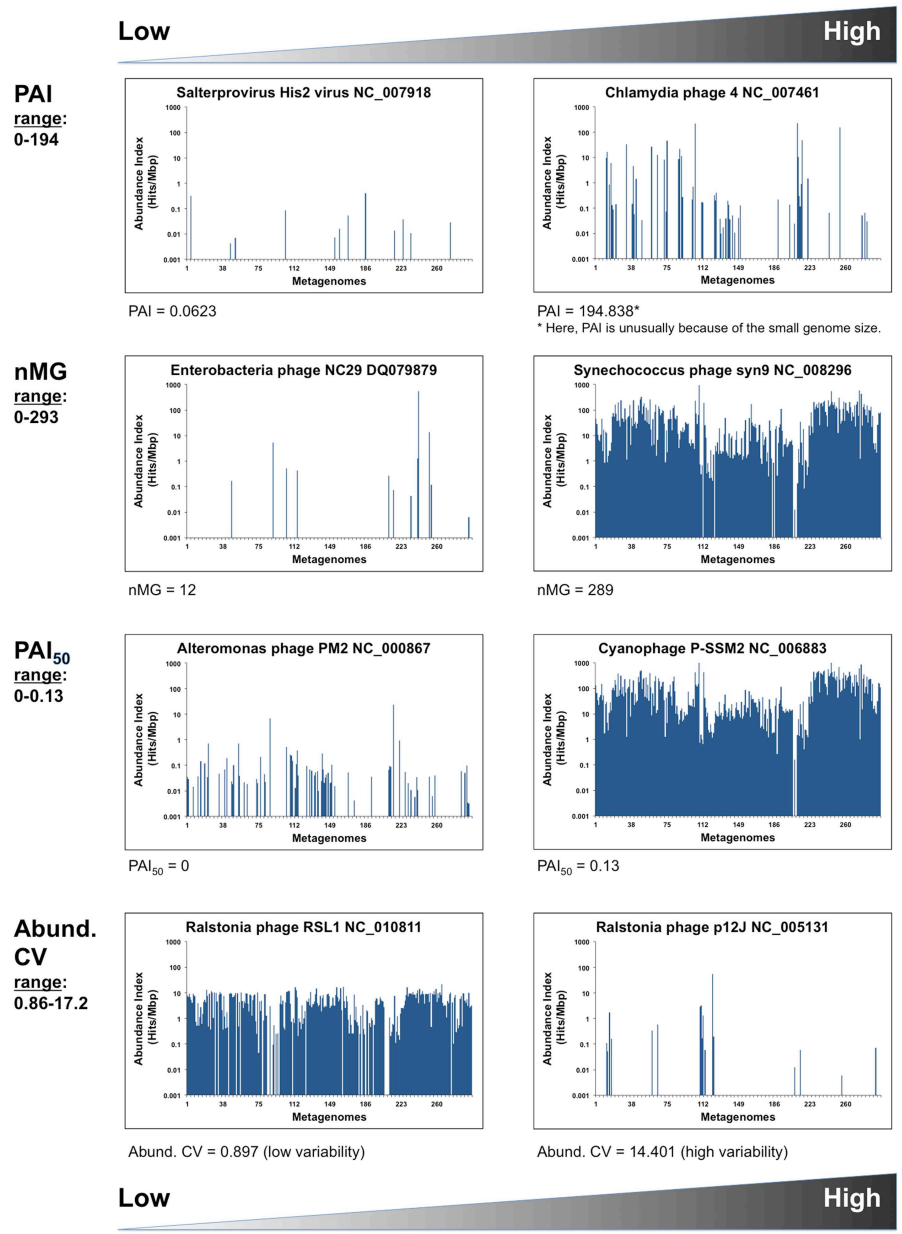


FIGURE 1 | Phage distribution metrics. Inter-phage metrics and statistics quantifying different aspects of phage abundance and distribution in 296 metagenomic samples. Graphical examples show the phage genomes at the high and low ends of each parameter. X-axes represent the metagenomes (MG) listed in the same order as in **Table S1** (i.e., grouped by environment). Y-axes are in logarithmic scales.

optimization of significance and alignment length thresholds). This value does not reflect the regularity or uniformity of the distribution, but indicates coverage gaps [sometimes referred to as metagenomic islands (Pasic et al., 2009; Mizuno et al., 2014)]. Other estimators of uniformity implemented in this study include the *spread* of a coverage plot (expressed as the coefficient of variation of coverage), *kurtosis* (a statistical value of a plot's uniformity), and an *adapted Shannon Evenness Index* applied to phage genes (explained in detail in **Table 3**). Examples of phage distribution and phage recruitment plots are provided in **Figures 1**, **2**, and all raw data are provided in **Table S2**.

Statistical Analysis

For statistical analysis, DataDesk (Data Description Inc., Ithaca, NY; URL: http://www.datadesk.com) and the R software environment (URL: http://www.r-project.org) were used.

Results

Input Data

Eco-Locator plots were generated for a core data set of 588 viral genomes and 296 metagenomes. Fragment-recruitment and coverage-density plots for each unassembled metagenome

TABLE 3 | Metrics used to describe phage ecological features at the nucleotide level.

Parameter	Definition/Calculation	Range	Significance/Interpretation/Limitations
PHAGE COVERAGE METRICS (INTRA-PHAGE OR NUCLEOTIDE-LEVEL METRICS): FOR A GIVEN RECRUITMENT PLOT OF A PHAGE X			
Coverage density (AUC/nNuc)	Area of a genome coverage plot (area-under-the curve) normalized to the total number of nucleotides in the phage genome.	0–2.920	This value is similar to the total abundance of a phage in all metagenomes; however, it also considers each nucleotide covered in the phage genome and not just the number of sequence reads that match that genome.
Density per metagenome (cumulative AUC/nMG)	Average overall phage density divided by the number of metagenomes.	126–1.71 $\times$ 10^6	This value normalizes the coverage density to the number of metagenomes in which the phage genome is found. It differentiates between the densities of ubiquitous phages (high nMG) and that of habitat-specific phages (low nMG).
%genome covered	Fraction of the phage genome that matches at least one metagenomic sequence.	0–100%	This value reflects the homogeneity of overall phage coverage in metagenomes as well as the gaps in coverage. It marks areas within a phage genome that have not been matched in any metagenomic sample, but is magnitude-independent—thus does not show which areas of the genome are overrepresented. A %genome coverage of 40% means that combined uncovered gaps are 60%.
Gene coverage evenness	Adapted Shannon Evenness Index (Shannon E) of the coverage of phage genes. $E = -\Sigma\ p_i \ln p_i$ /nGenes where p_i is the proportion of hits to the i^{th} gene to the sum of hits to all genes of phage X	0–0.92	This value reflects whether protein-encoding genes within a phage genome are equally represented relative to each other. A gene evenness of one means that all phage genes are equally represented (regardless of the magnitude of their coverage), while low evenness values suggest possible non-specific or cross-matching genes (i.e., parts or all of the phage genome is absent).
Coverage coefficient of variation (CV)	Standard deviation of coverage density/Mean coverage density (Coverage density = AUC/nNuc)	0.76–12.58	This value reflects the variation or spread of coverage along a phage genome. Typically a phage genome coverage plot with high CV has higher coverage values for certain parts of the genome and zero values for other parts.
Median coverage density	Median number of hits per nucleotide per phage	0–686	Less sensitive to extreme values, the median coverage density provides another indicator of the homogeneity of phage genome coverage in metagenomic samples.
Coverage kurtosis	Kurtosis equation: $\frac{\sum(X-\mu)^4}{N\sigma^4} - 3$ where X is the value of each data point, μ is the sample mean, σ is the standard deviation, and N is the number of data points	0.02–423.12	Kurtosis is a statistical measure of uniformity or lack thereof within a frequency distribution curve. It is often used as a measure of skewness, bimodality, or "peakiness" of a distribution plot. It has been adopted here to reflect the irregularity of a phage coverage density plot. If a phage genome coverage plot has high kurtosis, this means that some areas of this genome have sharp coverage peaks while others have low or no coverage values. Negative kurtosis values reflect flatter coverage plots but do not provide information about the coverage magnitude.

Representative examples of each value are given in ***Figure 2****.*

were generated and are publicly available (URL: http://www.phantome.org/eco-locator).

Implementation and Testing of Metagenome-Level Metrics

Abundance values (expressed as total AIs) of sequences related to known phages showed an immense variation among different metagenomes, spanning several orders of magnitude (range = 4–28,859 hits /Mbp; mean = 1462.8 hits /Mbp; median = 1125 hits /Mbp). At the lower end, samples from human lungs, classically thought to be free of resident microbiota, had the smallest fraction of sequences similar to known phages and the lowest sequence diversity and richness as previously reported (Willner et al., 2009, 2012) (**Table 4** and **Table S1**). Hypersaline samples also had low abundance indices, possibly resulting from the low number of completely sequenced viral sequences from these habitats (**Table S1**). At the other extreme, aquatic samples (both virus-enriched and microbial) contained the largest fraction of sequences similar to known phages. The microbial metagenome with highest phage AI was from the open ocean (Hydrostation S, Sargasso Sea, Bermuda), while the viral metagenome with highest phage AI was an estuary sample (Station 834, Chesapeake Bay Virioplankton) (**Table 4** and **Table S1**). The sample with

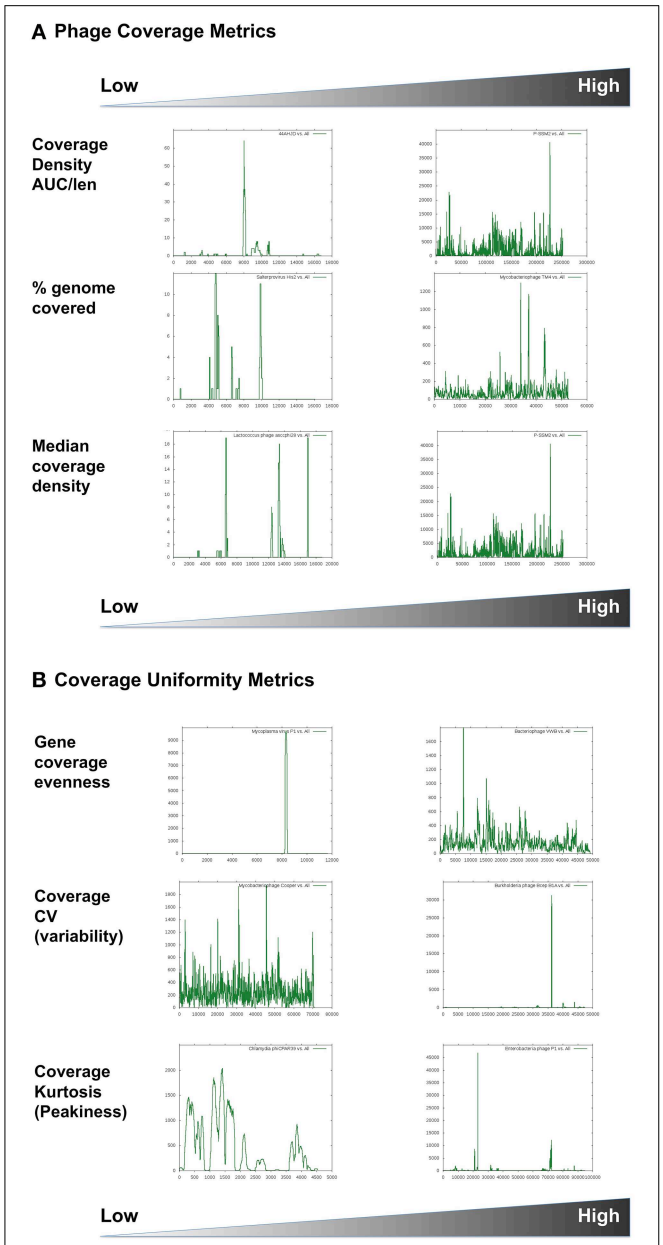


FIGURE 2 | Phage coverage metrics, including (A) density and (B) uniformity estimates. Graphical examples show high and low ends of each parameter used. X-axes represent the genome coordinates while Y-axes represent number of hits to each nucleotide. Graphs are scaled differently. The coverage plots are for the following phages: **(A)** *Staphylococcus* phage 44AHJD compared to Cyanophage P-SSM2; Salterprovirus His2 virus compared to Mycobacteriophage TM4; *Lactococcus* phage asccphi28 compared to Cyanophage P-SSM2. **(B)** *Mycoplasma* virus P1 compared to Bacteriophage VWB; Mycobacterophage Cooper compared to *Burkholderia cenocepacia* phage BcepB1A; *Chlamydia* phage phiCPAR39 compared to Enterobacteria phage P1.

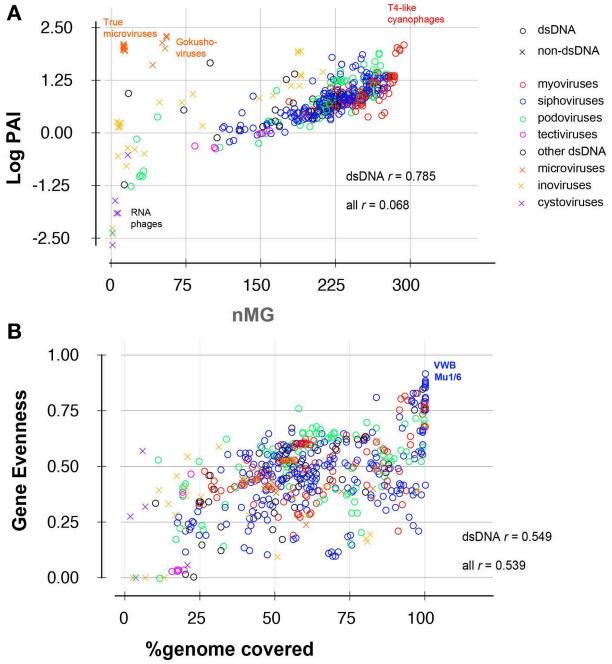


FIGURE 3 | Scatter plots showing correlation between (A) abundance and ubiquity or (B) gene evenness and % genome coverage of 588 viruses in 296 metagenomes. Data points are labeled according to phage family (different colors), and nucleic acid content (circles: dsDNA phages; crosses: other phages, i.e., ssRNA, dsRNA, and ssDNA phages). Correlation coefficients (*r*) are shown for all phages and for dsDNA phages alone.

highest number of phage types (richness) was from a marine-derived lake in Antarctica, and those with highest phage sequence diversity (Shannon diversity) were human gut samples (**Table 4**).

An in-depth ecological analysis comparing all metagenomes or examining phage habitat-association is beyond the scope of this Methods Article; however, a glimpse at extreme values of each metric (**Table 4**) provides confidence in the methodology used because of its agreement with previous analyses performed on subsets of those data (Angly et al., 2006; Dinsdale et al., 2008; Willner et al., 2009) and because of some biologically relevant measurements (such as the low phage richness in lungs or the high phage diversity in stool samples).

Implementation and Testing of Phage-Level Metrics

The most common statistics used in viral metagenomic studies rely on two key parameters: the relative abundance of phage-like sequences [defined here as PAI and referred to as depth in some other studies (Dutilh et al., 2014; Martinez Martinez et al., 2014)] and the nMGs in which a particular phage is represented (ubiquity or nMG) (e.g., Mizuno et al., 2013; Dutilh et al., 2014). These two statistics are undoubtedly useful, but are limited by the following: (i) we observed that plotting abundance and ubiquity successfully resolves classes of RNA or single-stranded DNA (ssDNA) viruses, yet these two metrics are partly interdependent among double-stranded DNA viruses (correlation index = 0.785, **Figure 3A**); (ii) abundance and ubiquity metrics quantitatively describe phage prevalence but do not describe the *pattern* of this prevalence (e.g., phage-ecosystem correlations or

TABLE 4 | Examples of the lowest and highest scoring metagenomes or phages according to different metrics.

Parameter	Low	High
METAGENOME-LEVEL METRICS:		
Total AI	Lung samples (**Table S1**) (Values: 4.07–8.22)	Hydrostation S, Sargasso Sea, Bermuda (open ocean) (Value = 28.859)
Median AI (AI_{50})	Lung samples (**Table S1**) (Value = 0)	Chesapeake Bay, MD (estuary): Chesapeake Bay Virioplankton–Station 834 (Value = 3.061)
nPhages	Viral data from the human lung (Sample 109) Value = 8 phages	AntarcticaAquatic_5–Marine-derived lake (Value = 487 phages)
Shannon Diversity Index	Viral data from the human lung (Sample 109) Value = 2.061	Stool metagenome (sample 179) Value = 5.813
Shannon evenness E	GS051 Shotgun–Coral Reef Atoll–Polynesia Archipelagos–Rangirora Atoll–Fr. Polynesia (Value = 0.008)	Viral data from the human lung (sample 109) Value = 0.258
PHAGE DISTRIBUTION METRICS:		
Phage abundance index (PAI)	Eleven out of 17 RNA phages have zero values	*Chlamydia* phage 4 (ssDNA) Value = 194.84; Cyanophage P-SSM4 (dsDNA) Value = 109.856
PAI_{50}	*Aeromonas* phage PM2 (Value = 0)	T4-like cyanophage P-SSM2 (Value = 0.13)
nMG	Eleven RNA viruses have zero values; *Pseudomonas* phi-6 (dsRNA, Value = 1); dsDNA: *Lactococcus* phage asccphi28 (Value = 20)	T4-like cyanophage P-SSM2 (Value = 293)
Abund. CV	Myoviridae Bacillus phage 0305phi8-36 (Value = 0.86)	Ralstonia phage P12 J (dsDNA, Value = 14.4), *Pseudomonas* phage phi-6 (dsRNA, Value = 17.2) and microviruses (ssDNA, Values > 16)
WITHIN PHAGE COVERAGE/DENSITY METRICS (INTRAPHAGE PROPERTIES):		
Coverage density	Levivirus Enterobacteria phage MS2 (ssRNA, Value = 0.04); *Staphylococcus* phage 44AHJD (dsDNA, Value = 1.03)	Coliphage phiX174 (ssDNA, Value = 2.920); T4-like cyanophage P-SSM2 (1.989)
Density per metagenome	Enterobacteria phage MS2 (ssDNA, Value = 126); *Lactococcus* phage asccphi28 (dsDNA, Value = 540.45)	T4-like cyanophage P-SSM2 (1.71×10^6)
%genome covered	Salterprovirus His 2 (Value = 10%; lowest non-zero value for a dsDNA virus)	Mycobacteriophages Rosebush and Cooper (Value = 100%)
Gene coverage evenness	*Mycoplasma* virus P1 (lowest non-zero value = 0.003)	Bacteriophage VWB (Value = 0.918) and *Streptomyces* Mu1/6 (Value = 0.886)
Spread (CV)	Actinoplanes phage phiAsp (Value = 0.757)	*Burkholderia* phage BcepB1A (Value = 12.581)
Coverage kurtosis	*Chlamydia* phage phiCPAR39 (ssDNA, Value = 0.02); unclassified Picovirinae Actinomyces phage Av-1 (dsDNA, Value = 3.03)	Enterobacteria phage P1 (Value = 423.12)
Median density	*Lactococcus* phage Asccphi28 (among 254 phages with zero value)	T4-like cyanophage P-SSM2 (Value = 686)

If the high end is not a dsDNA phage, the next highest/lowest dsDNA phage is also shown.

habitat-specificity); (iii) these values are sensitive to biases (for example, they may be strongly affected by the dominance of aquatic samples or human-associated samples in a data set). Accordingly, we implemented additional metrics to better assess the multidimensional nature of abundance and distribution of phage sequences as well as the intra-phage coverage density and evenness (detailed in Methods and **Tables 2**, **3**). For example, we estimated the cross-habitat variation among AIs using the coefficient of variation (Abundance CV; **Tables 2** and **Table S2**), which provides information on the homogeneity of distribution of phage sequences across metagenomes, and can differentiate between cosmopolitan and habitat-confined phages (Thurber, 2009).

Eleven RNA phages in our database were practically undetected. The absence of these RNA viruses is expected since the metagenomes analyzed consisted only of DNA and were not supposed to contain RNA contamination, and since there is little shared sequence similarity between RNA and DNA phage genes, as seen in the Phage Proteomic Tree (Rohwer and Edwards, 2002) and the Phage Population Network (Lima-Mendez et al., 2008).

Another important set of metrics implemented in this study describe the uniformity of sequence coverage within a phage genome, and thus help indicate whether phage abundance values represent presence of an entire related phage or result from the overabundance of specific conserved genes or tiny fractions of phage genomes. Of those values, the % sequence coverage in all metagenomes, for example, gives a good indication of the global distribution of phage modules, while the gene evenness parameter is an indicator of the covariation of different genes between different habitats (**Figures 2A**, **3B**). Overall, more than a dozen metrics were used to describe the ecological and coverage properties of each phage genome (**Table S2**), 11 of which were selected (**Tables 1**, **2**) and combined to separate all phage genomes based on two principal components that summarize the 11 dimensions and explain ∼65% of the variance (**Figure 4**).

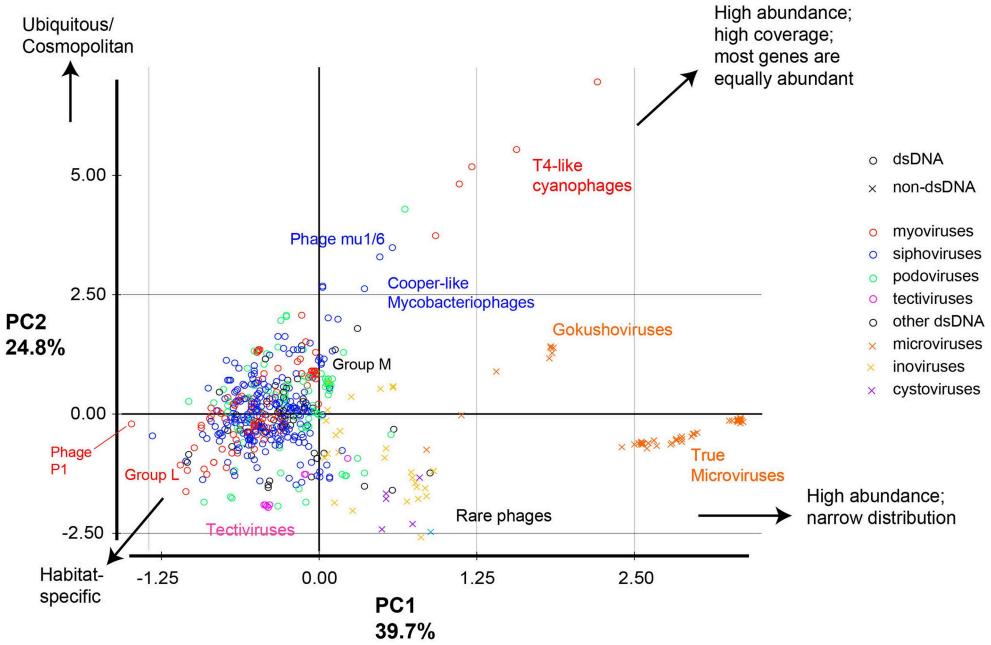


FIGURE 4 | Principal component analysis of phage genomes according to their ecological properties. All phages were compared based on 11 metrics, then the 11 dimensions were reduced into two principal components that explain most of the variance. Circles represent dsDNA phages and x signs represent other types of phage genomes; colors represent different phage classes. Examples of phages and groups of phage discussed in the text are labeled.

Combining the Multidimensional Metrics Separates Phage Genomes Based on Ecological Parameters

Taken together, this combination of metrics allowed the separation of phages into distinct groups (**Figure 4**) according to their environmental abundance, distribution, and sequence coverage parameters. The most prominent groups are:

- Phages with high abundance, broad distribution, and low inter-sample variation (e.g., T4-like cyanophages, *Bordetella* phages, and *Streptomyces* Phage Mu1/6). This pattern indicates ubiquitous or near cosmopolitan phages, or—alternatively—phages highly similar to cosmopolitan phages.
- Phages with high abundance, broad distribution, but low gene evenness (e.g., Cooper-like mycobacteriophages).
- Phages with high abundance and narrow distribution (high coefficients of variation between metagenomes). This pattern means very high abundance in only a few metagenomes, but partial genome representation. This group mostly consists of the ssDNA *Microviridae* and is further divided into the gokushoviruses (such as the *Chlamydia* phages, which had relatively high percent sequence coverage) and the true microviruses (such as phiX174, which had low percent sequence coverage) (Labonte and Suttle, 2013).
- Phages with low abundance but wide distribution (e.g., some *Pseudomonas* phages (phiKZ, phage 201phi 2-1, and phage EL). This pattern suggests a wide distribution of some highly conserved genes or modules within those phages (Group L in **Figure 4**).
- Phages with low abundance but high percent sequence coverage (e.g., *Pseudomonas* phage MP38, *Pseudomonas* phage MP29, *Pseudomonas* phage MP22, and Bacteriophage D13 112). These are referred to as Group M in **Figure 4**.
- Rare phages (e.g., some *Vibrio* phages of the ssDNA phage class *Inoviridae* such as: phages VEJphi, VGJphi, VSK, KSF-1phi, and O139 fs1).

In summary, the metrics were particularly useful in determining outliers or extreme phage groups (e.g., microviruses, cyanoviruses, tectiviruses, etc…). The analysis highlighted the scarcity of sequences shared with RNA phages, the massive yet uneven observed abundance of microviral sequences, and the dominance of T4-like cyanophages and Cooper-like mycobacteriophages in currently sequenced metagenomes.

Discussion

Estimating phage diversity in nature has generally been more difficult than estimating the diversity of cellular microorganisms—whether by culture-based or molecular methods. This difficulty is, in part, caused by the lack of a set of universal genes common to all phages that can be used for phylogenetic profiling, as opposed to ribosomal DNA and tRNA synthetase genes in cellular life forms (i.e., domains: Archaea, Eubacteria, and Eukartyota). Thus, the emergence of metagenomics has been particularly useful for phage biologists by providing a method for surveying complete phage communities (Breitbart et al., 2002; Angly et al., 2005, 2009; Edwards and

Rohwer, 2005). One particularly interesting aspect of these analyses has been the realization that the majority of viral metagenomic sequences do not have any similarities to the databases, highlighting the large amount of "viral dark matter" in the universe. However, the distribution of sequences similar to completely sequenced phage genomes provides important information about the distribution of these representative, well-characterized phages in natural systems. Whereas, early metagenomic studies were highly descriptive in nature, the phenomenal accumulation of metagenomic data now enables researchers to advance from cataloging phage species and functional categories to addressing fundamental questions about phage ecology, evolution, and phage-host co-occurrence and co-evolution. Such questions require the establishment of methods and metrics beyond simply counting metagenomic sequence reads recruited to a phage or taxonomic binning.

In this study, we expand the available analyses for examining phage distributions in unassembled metagenomes by adapting metrics to quantify not only fragment and gene counts, but also (i) coverage density, depth, uniformity, and breadth of phage sequence distribution in metagenomic data sets; and (ii) extent of variability of sequence recruitment to a given phage genomic scaffold. These metrics allowed us to separate phages into groups that more accurately reflect their ecology, which will allow the examination of phage-habitat and phage-host associations in future studies as a wider range of metagenomes are sequenced.

The present work did not aim at developing novel statistical functions or mathematical equations, but rather adapted well-established functions and, sometimes, repurposed metrics used in other fields or applications (such as evenness and kurtosis). The following attributes distinguish the set of metrics that were implemented:

- ***Multiple-level normalization:*** Counting sequence similarity hits is probably the most straightforward and most popular indicator of the abundance of genes and genomes in an ecosystem. With the availability of multiple data sets with different sequence depths and variable read lengths, it has become common practice to normalize the number of hits to the metagenome size (expressed as the number of reads or preferably in as the number of base pairs). Moreover, since a metagenomic data set is just a sample of all the DNA in an environment, any gene (or genome) is more likely to be represented in that sample if it is: (i) more abundant or (ii) larger in size (number of base pairs). Thus, we also normalized hit counts to the length of the gene or genome to which they recruited. The concept of length-normalization is often used in RNA-Seq analysis (Lee et al., 2011) and was introduced by Angly and coworkers in the GAAS suite for estimating relative abundances of full-length phages (Angly et al., 2009). Here, we adopted and expanded length normalization for every analyzed entity (whether it's a protein-encoding gene, genome, or a genomic fragment).
- ***Estimation of coverage density and uniformity:*** Because phage genomes are known for their high mosaicism and because they often contain protein-encoding genes with a wide range of conservation and so-called metagenomic islands (Pasic et al., 2009; Mizuno et al., 2014), we deployed metrics to assess the uniformity vs. variability patterns of coverage plots. For this we describe three different parameters: (i) density or depth, (ii) uniformity or evenness, and (iii) regularity or peakiness. To measure density, we adopted the commonly used measure of number of hits per nucleotide, or the normalized area under the curve (AUC/nNuc) of a coverage plot. To describe coverage uniformity, we used both the coefficient of variation (CV) as an estimator of the *spread* of a coverage plot and the Shannon Evenness metric (E) as an estimator of gene coverage *evenness* in a given genome. Finally, we adopted the *kurtosis* metric that is used to describe distribution curves or line graphs as an estimator of the regularity/irregularity of peaks in a coverage plot.
- ***Multidimensional analysis.*** Each of the developed metrics utilized has different strengths and weaknesses. Under specific conditions, some metrics may be more informative than others; some of them may partly correlate; and some could be redundant in certain conditions (e.g., highly abundant and uniformly covered phage genomes will have similar median coverage density and evenness). To take advantage of all the information provided by the different metrics without being misled by one or two of them, we used PCA analysis, which effectively split phages into groups reflecting both their sequence similarity and their ecological distribution.

Potential Limitations and Suggested Solutions

For some specific phage groups, such as T4-like phages and microviruses, assigning a phage genome was quite difficult. For example, the apparent prevalence of non-marine T4-like phages in most samples may be a result of the overabundance of their closely related cyanophage T4-like genes. In support of this interpretation is the observation that the distribution pattern of phage T4 genome was overshadowed by that of the T4-like cyanophage, P-SSM2 (Ignacio-Espinoza and Sullivan, 2012), especially in ecosystems in which T4-like cyanophages were abundant. In such cases, coverage metrics are crucial in determining whether an entire phage is present in a particular ecosystem, or if the distribution more likely results from conserved genes.

A more striking example is ssDNA phages. Microviruses are ssDNA phages that have previously been shown to be quite abundant in certain metagenomes, especially those created using rolling circle amplification with the phi29 polymerase (e.g., Desnues et al., 2008; Lopez-Bueno et al., 2009; Tucker et al., 2011). Currently sequenced *Microviridae* include the gokushoviruses, which infect obligate intracellular parasites such as *Chlamydia, Spiroplasma*, and *Bdellovibrio*, and the true microviruses (such as phiX174) that infect enteric bacteria (Labonte and Suttle, 2013). However, examining the coverage patterns reveals that most metagenomic sequence reads that match the true microviruses are similar to a tiny fraction of the genome, while the gokushoviruses are frequently covered at nearly 70% (**Figure 2B**). This pattern of coverage suggests that ssDNA viruses similar to the gokushoviruses are present in the environments examined, while the true microviruses

are likely not present. This is an important distinction since simple measurements of abundances would likely miss that distinction, suggesting an abundance of both groups. Another important revelation of this analysis is the confirmation that microviruses were only identified in a limited nMGs, which were amplified using phi29 polymerase, which is known to disproportionately amplify small, circular, ssDNA genomes (Kim and Bae, 2011). However, since the methods used for constructing and sequencing the other metagenomes may have excluded ssDNA viruses, the actual presence or abundance of gokushoviruses in other environments remains unknown. In either case, the relative abundance of these genomes, in particular, is not thought to reflect their natural occurrence.

Finally, in the data sets described here (**Table S2**), most phages had less than 75% overall sequence coverage per genome (68% of dsDNA phages and 85% of non dsDNA phages were <75% covered). While sequencing depth is a major factor controlling coverage—especially in the case of rare phages, another reason behind this low coverage is that sampled phage genomes may be only partly similar to those in databases while they have other unique, yet-to-be-sequenced modules. This is a limitation that can be addressed through assembling metagenomes, and will likely be reduced as more phages are sequenced and publicly deposited.

Portability and Reproducibility of the Methods

The metrics described above are intended to be platform-independent, i.e., they can be applied to any metagenomic analysis pipeline that generates recruitment plots or that map metagenomic hits to a sequence contig/scaffold. The metagenome-level metrics (**Table 1**) and inter-phage metrics (**Table 2**) can be applied to any metagenome vs. phage data matrix, where the number of metagenomic reads per phage is calculated at a given *E*-value threshold. The coverage metrics (intra-phage properties, **Table 3**) can be generated from any recruitment plot where metagenomic sequences are mapped to a phage scaffold or contig. Although we used tBlastX output for mapping, we believe that any other similarity search or mapping tool can be used as well.

Of course, the key to reproducibility in any such analysis is to use the same database/reference set for all comparisons, i.e., the same set of phage genomes has to be used for analyzing all metagenomic data sets, if the results are to be compared to one another. If more phage genomes are added to the Blast database, for example, then any older analyses have to be repeated against the updated database. This is true for any (meta)genomic annotation or analysis pipeline.

Conclusion

In conclusion, we expanded the existing repertoire of viral metagenomic analysis tools by implementing an array of metrics to describe different aspects of the ecological distribution of archaeal viruses, phages and phage-like sequences in metagenomic data sets. Some of these metrics have been well-developed and efficiently used in phage metagenomic bioinformatics, while others have been used for the first time in this study or adopted from other mathematical and statistical applications and repurposed toward phage analysis. Together, this suite of metrics is useful in expressing different dimensions of phage abundance, extent and breadth of distribution, as well as phage sequence coverage depth and uniformity in diverse ecosystems. The combination of these metrics successfully separates phages in ecologically meaningful ways, which will enable researchers to generate and test biological hypotheses regarding phage ecology and evolution.

Author Contributions

Conceived and designed the study: RA, MB, RE. Developed and applied the methods: RA. Designed tools and wrote scripts: RA, SA, RE. Performed the experiments: RA, BD, SA, RE. Analyzed the data: RA, BD, SA, MB. Wrote the paper: RA, BD, MB, RE.

Funding

This work was supported by the PhAnToMe grant from the National Science Foundation Division of Biological Infrastructure (DBI-0850356 to RE and DBI-0850206 to MB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank Robert Schmieder and Nick Celms for technical help with Perl. The web tool used in this article, Phage Eco-Locator, was presented at the UT-ORNL-KBRIN 2011 Bioinformatics Summit, Memphis-TN, and the presentation abstract was published as part of the conference proceedings (Aziz et al., 2011).

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00381/abstract

Figure S1 | Summary statistics of the phage genomes and metagenomic libraries used. (A) Metagenomic samples classified and sorted by their environments; **(B)** Phages classified by the bacterial families they infect; **(C)** Phages grouped into taxonomic classes.

Table S1 | Raw data calculated for each metagenomic sample used in the study, together with relevant metadata for each metagenome (e.g., biome, environment, size in base pairs, number of reads, GC%).

Table S2 | Raw data for all phages, their metadata, and all analysis metrics applied to them.

Both **Tables S1** and **S2** are also available online (URL: http://www.phantome.org/eco-locator/v1_tables). The online tables are updatable (e.g., in the case of a change of phage name or the availability of more accurate metagenomic metadata).

References

Abedon, S. T. (2009). Phage evolution and ecology. *Adv. Appl. Microbiol.* 67, 1–45. doi: 10.1016/S0065-2164(08)01001-0

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

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

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Angly, F. E., Willner, D., Prieto-Davo, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., et al. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.* 5:e1000593. doi: 10.1371/journal.pcbi.1000593

Aziz, R., Dwivedi, B., Breitbart, M., and Edwards, R. (2011). Phage Eco-Locator: a web tool for visualization and analysis of phage genomes in metagenomic data sets. *BMC Bioinformatics* 12:A9. doi: 10.1186/1471-2105-12-S7-A9

Belcaid, M., Bergeron, A., and Poisson, G. (2010). Mosaic graphs and comparative genomics in phage communities. *J. Comput. Biol.* 17, 1315–1326. doi: 10.1089/cmb.2010.0108

Bergh, O., Børsheim, K. Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature* 340, 467–468. doi: 10.1038/340467a0

Breitbart, M. (2012). Marine viruses: truth or dare. *Ann. Rev. Mar. Sci.* 4, 425–448. doi: 10.1146/annurev-marine-120709-142805

Breitbart, M., Hewson, I., Felts, B., Mahaffy, J. M., Nulton, J., Salamon, P., et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* 185, 6220–6223. doi: 10.1128/JB.185.20.6220-6223.2003

Breitbart, M., and Rohwer, F. (2005). Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 39, 729–736. doi: 10.2144/000112019

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Caporaso, J. G., Knight, R., and Kelley, S. T. (2011). Host-associated and free-living phage communities differ profoundly in phylogenetic composition. *PLoS ONE* 6:e16900. doi: 10.1371/journal.pone.0016900

Desnues, C., Rodriguez-Brito, B., Rayhawk, S., Kelley, S., Tran, T., Haynes, M., et al. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature* 452, 340–343. doi: 10.1038/nature06735

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., et al. (2008). Functional metagenomic profiling of nine biomes. *Nature* 452, 629–632. doi: 10.1038/nature06810

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 4498. doi: 10.1038/ncomms5498

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

Ignacio-Espinoza, J. C., and Sullivan, M. B. (2012). Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. *Environ. Microbiol.* 14, 2113–2126. doi: 10.1111/j.1462-2920.2012.02704.x

Kim, K. H., and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77, 7663–7668. doi: 10.1128/AEM.00289-11

Labonte, J. M., and Suttle, C. A. (2013). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Front. Microbiol.* 4:404. doi: 10.3389/fmicb.2013.00404

Lee, S., Seo, C. H., Lim, B., Yang, J. O., Oh, J., Kim, M., et al. (2011). Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.* 39, e9. doi: 10.1093/nar/gkq1015

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* 25, 762–777. doi: 10.1093/molbev/msn023

Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., and Alcami, A. (2009). High diversity of the viral community from an Antarctic lake. *Science* 326, 858–861. doi: 10.1126/science.1179287

Martinez Martinez, J., Swan, B. K., and Wilson, W. H. (2014). Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J.* 8, 1079–1088. doi: 10.1038/ismej.2013.214

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Mizuno, C. M., Ghai, R., and Rodriguez-Valera, F. (2014). Evidence for metaviromic islands in marine phages. *Front. Microbiol.* 5:27. doi: 10.3389/fmicb.2014.00027

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013). Expanding the marine virosphere using metagenomics. *PLoS Genet.* 9:e1003987. doi: 10.1371/journal.pgen.1003987

Pasic, L., Rodriguez-Mueller, B., Martin-Cuadrado, A. B., Mira, A., Rohwer, F., and Rodriguez-Valera, F. (2009). Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics* 10:570. doi: 10.1186/1471-2164-10-570

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., et al. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J.* 4, 739–751. doi: 10.1038/ismej.2010.1

Rohwer, F., and Edwards, R. (2002). The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol.* 184, 4529–4535. doi: 10.1128/JB.184.16.4529-4535.2002

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., et al. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* 27, 3074–3075. doi: 10.1093/bioinformatics/btr519

Schmieder, R., and Edwards, R. (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288

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

Schmieder, R., Lim, Y. W., Rohwer, F., and Edwards, R. (2010). TagCleaner: identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* 11:341. doi: 10.1186/1471-2105-11-341

Shannon, C. E. (1948). A Mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423; 623–656. doi: 10.1002/j.1538-7305.1948.tb01338.x

Silva, G. G., Cuevas, D. A., Dutilh, B. E., and Edwards, R. A. (2014). FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2:e425. doi: 10.7717/peerj.425

Spellerberg, I. F., and Fedor, P. J. (2003). A tribute to Claude Shannon (1916-ì2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Glob. Ecol. Biogeogr.* 12, 177–179. doi: 10.1046/j.1466-822X.2003.00015.x

Swanson, K. S., Dowd, S. E., Suchodolski, J. S., Middelbos, I. S., Vester, B. M., Barry, K. A., et al. (2011). Phylogenetic and gene-centric metagenomics of the canine intestinal microbiome reveals similarities with humans and mice. *ISME J.* 5, 639–649. doi: 10.1038/ismej.2010.162

Thurber, R. V. (2009). Current insights into phage biodiversity and biogeography. *Curr. Opin. Microbiol.* 12, 582–587. doi: 10.1016/j.mib.2009.08.008

Thurber, R. V., Willner-Hall, D., Rodriguez-Mueller, B., Desnues, C., Edwards, R. A., Angly, F., et al. (2009). Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* 11, 2148–2163. doi: 10.1111/j.1462-2920.2009.01935.x

Tucker, K. P., Parsons, R., Symonds, E. M., and Breitbart, M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J.* 5, 822–830. doi: 10.1038/ismej.2010.188

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* 28, 127–181. doi: 10.1016/j.femsre.2003.08.001

Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proc. Natl. Acad. Sci. U.S.A.* 95, 6578–6583. doi: 10.1073/pnas.95.12.6578

Williamson, K. E., Radosevich, M., and Wommack, K. E. (2005). Abundance and diversity of viruses in six Delaware soils. *Appl. Environ. Microbiol.* 71, 3119–3125. doi: 10.1128/AEM.71.6.3119-3125.2005

Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., Silva, J., et al. (2009). Metagenomic analysis of respiratory tract DNA viral communities in

cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 4:e7370. doi: 10.1371/journal.pone.0007370

Willner, D., Haynes, M. R., Furlan, M., Schmieder, R., Lim, Y. W., Rainey, P. B., et al. (2012). Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J.* 6, 471–474. doi: 10.1038/ismej.2011.104

Wommack, K. E., and Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* 64, 69–114. doi: 10.1128/MMBR.64.1.69-114.2000

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., et al. (2013). Abundant SAR11 viruses in the ocean. *Nature* 494, 357–360. doi: 10.1038/nature11921

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

frontiers in
MICROBIOLOGY

PERSPECTIVE ARTICLE
published: 07 January 2015
doi: 10.3389/fmicb.2014.00739

Finding and identifying the viral needle in the metagenomic haystack: trends and challenges

Hayssam Soueidan[1,2], Louise-Amélie Schmitt[1,3], Thierry Candresse[4,5] and Macha Nikolski[1,3]*

[1] *Bordeaux Bioinformatics Center, Université de Bordeaux, Bordeaux, France*
[2] *INSERM U1035, Université de Bordeaux, Bordeaux, France*
[3] *Centre National de la Recherche Scientifique/Laboratoire Bordelais de Recherche en Informatique, Université de Bordeaux, Talence, France*
[4] *Institut National de la Recherche Agronomique, UMR 1332 Biologie du Fruit et Pathologie, Villenave d'Ornon, France*
[5] *UMR 1332 Biologie du Fruit et Pathologie, Université de Bordeaux, Villenave d'Ornon, France*

Edited by:
Bas E. Dutilh, Radboud University Medical Center, Netherlands

Reviewed by:
Ivan Merelli, Institute for Biomedical Technologies, Italy
Simon Roux, University of Arizona, USA

***Correspondence:**
Macha Nikolski, Bordeaux Bioinformatics Center, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux, Bordeaux, France
e-mail: macha.nikolski@labri.fr

Collectively, viruses have the greatest genetic diversity on Earth, occupy extremely varied niches and are likely able to infect all living organisms. Viral infections are an important issue for human health and cause considerable economic losses when agriculturally important crops or husbandry animals are infected. The advent of metagenomics has provided a precious tool to study viruses by sampling them in natural environments and identifying the genomic composition of a sample. However, reaching a clear recognition and taxonomic assignment of the identified viruses has been hampered by the computational difficulty of these problems. In this perspective paper we examine the trends in current research for the identification of viral sequences in a metagenomic sample, pinpoint the intrinsic computational difficulties for the identification of novel viral sequences within metagenomic samples, and suggest possible avenues to overcome them.

Keywords: microbial metagenomics, NGS, virome, host—pathogen interactions, taxonomic assignment

INTRODUCTION

While genomics is the research field relative to the study of the genome of any organism, metagenomics is the term coined for the research that focuses on many genomes at the same time, as typical in some sections of environmental studies. The analysis of microbial communities has been until recently a complicated if not untractable task due to their high diversity and to the fact that many of these organisms cannot be cultured. Harnessing the major advances achieved in sequencing technologies, metagenomics has emerged as the only currently available approach to extensively characterize these largely unculturable communities. Besides vastly enriching our knowledge of microbial diversity in a varied range of environments, and providing information on the dynamics and on the overall functioning of microbial communities, metagenomics is also shedding light on many important biological processes and, in particular, on the role of the microbiome in biological functions essential for the development of higher order organisms harboring it (Blottière et al., 2013; Manor et al., 2014), or in the development of pathological problems (Cénit et al., 2014; Vayssier-Taussat et al., 2014). In addition, metagenomic efforts also vastly enrich the repertoire of genes available for biotechnological applications (Ni and Tokuda, 2013).

At the same time metagenomics extensively relies on bioinformatics to tackle the huge amounts of sequence data involved, and recognizes the need to develop computational methods that maximize our understanding of the genetic composition and the biological activities expressed in communities so complex that they can only be sampled, never completely characterized. Computational analysis has become a genuine bottleneck for metagenomics due not only to the large amount of sequence data, but also to the new questions such as, for example, the need for simultaneous assembly of multiple genomes or transcriptomes and the analysis of complex networks of host-microbe interactions (Wooley and Yuzhen, 2009).

In this context the analysis of viral communities presents particular interest but also computational challenges. The ability to thoroughly analyze the viral composition of an environmental sample is of paramount importance, in particular because viruses have turned out to play a major role in the functioning of microbial communities by processes such as viral infection and selective killing of certain taxa or as vectors for horizontal gene transfer (Suttle, 2007). Consequently, the viral part of the microbiome has been shown in a number of situations to have a major impact on the dynamics and on the evolutionary processes of their host populations. The discovery and classification of novel viral species, but also of higher order taxa, is therefore of particular interest in this context (Rosario and Breitbart, 2011).

One of the main goals of metagenomic projects is to characterize the microbial communities in terms of the identity and diversity of species present (species richness) in a given environment. When it comes to species identification, the task is called taxonomic assignment. Current NGS technologies have provided an opportunity for doing this analysis routinely (Petrosino et al., 2009). Software tools for automated taxonomic assignment for organisms such as bacteria and fungi have since become a mature technology and are now routinely used in many studies.

If bacterial or fungal applications have recently seen major advances, the problem of taxonomic assignment for

viruses—such as it arises in environmental studies—remains largely unsolved from the computational point of view, as exemplified by the difficulty of distinguishing viral genomes from eukaryotes and bacteria observed in some studies (Bazinet and Cummings, 2012). Indeed, ab-initio identification of a sequence as belonging to a cellular organism or to a virus remains a complicated task outside of the popular sequence-homology based approaches that rely on direct comparisons with already known viral sequences present in international databases.

We distinguish the task of deciding to which first-level domain (eukaryotes, bacteria, archaea, virus) a given sequence belongs—that we call first-level assignment—from a more fine-grained taxonomic assignment at, e.g., family, genus or species level. In virome studies the latter task is greatly facilitated when targeted sequencing of purified viral particles is performed (Hall et al., 2014), but the former is particularly difficult for complex samples containing both eukaryotic and viral sequences and when, as is very frequently the case, unknown viral species are present. In this paper we examine reasons behind this difficulty and suggest possible avenues to overcome them.

FIRST-LEVEL CLASSIFICATION OF COMPLEX ENVIRONMENTAL SAMPLES

The first-level assignment of sequence data coming from a non-targeted sequencing of a metagenomic sample is a particularly challenging computational problem. The most blatant difficulty is in the recognition of novel viral sequences, for which no close homologs have been previously characterized. This question is however of paramount importance for the biologists. From the biodiversity point of view, the identification of unknown viruses representing novel higher order taxa (genera, families. . .) is of clear interest as evidenced, for example, by the discovery of the Mimiviruses with genomes exceeding in size those of many bacterial genomes (Claverie et al., 2009). But this question can also have important practical implications as when trying to identify novel viruses responsible for particular syndromes or diseases in humans, plants or husbandry animals (Roossinck, 2012; Lecuit and Eloit, 2014).

A number of bioinformatics methods efficiently perform the first-level assignment of sequences from a sample mainly containing known species. Computational solutions can be broadly organized in two main categories: (1) sequence similarity methods and (2) sequence composition methods.

Methods that rely on sequence similarity can be themselves subdivided in alignment-based techniques (mostly attempting to improve BLAST accuracy) and index-based. Alignment-based methods suffer from two limitations: speed and lack of sensitivity (e.g., Bazinet and Cummings, 2012; Wood and Salzberg, 2014). Recently, novel solutions have been suggested to overcome these limitations. These methods are based on long k-mers (words of size k) and conceptually rely on the fact that when k is sufficiently large, k-mers become very specific. Consequently, the idea is to index the databases by long k-mers. This is indeed the foundation of MegaBlast (a general-purpose sequence aligner using long seeds), but also of a number of methods specific for taxonomic assignment such as LMAT (Ames et al., 2013) and Kraken (Wood and Salzberg, 2014). The downside of these approaches is over-specificity, which makes classification of unknown sequences problematic. This limitation can be particularly acute given the known very high intraspecific variability existing in some viral species or higher order taxa. For example, current criteria of the International Committee for the taxonomy of viruses tolerate up to 28% of nucleotide sequence divergence for the polymerase or capsid protein genes for isolates of a same species in the *Betaflexiviridae* family and a similar level of divergence at the whole genome level in the *Potyviridae* family (King et al., 2012).

A complementary approach is based on sequence composition analysis. Such methods rely on the decomposition of sequences into frequencies of short k-mers and make use of machine learning techniques (e.g., SVM, kNN, Naive Bayes, etc.) to train a classifier on a reference database. The taxonomic assignment of novel sequences is then predicted by applying the pre-trained model. These methods theoretically are better suited to the task of novel species classification as short k-mers distributions are less prone to over-fitting. However, even these techniques fail to classify about 50% of species absent from the training set (Nalbantoglu et al., 2011). This is especially salient for viral sequences, as the vast majority of them fail to be uniquely assigned to any domain of life (Rosen et al., 2010).

Recent results show that *contig*-level assembly improves the strength of the taxonomic signal contained in individual short reads, even in the case of increased chimericity (Mende et al., 2012; Teeling and Glockner, 2012). This is why in our experimental evaluation (see Section Why is the First-level Assignment Problem Hard?) we work exclusively with sequence lengths that are comparable to contigs obtained by a standard metagenomic assembly step when the data originate from complex biological communities.

In summary, even the simple goal to provide a first-level description of a sample composition and be able to reveal if viral sequences are present, has been eluding a satisfactory solution. Indeed, for viral (and also eukaryotic) sequences, none of the existing methods produces a taxonomic distribution that is even remotely close to the expected one (Bazinet and Cummings, 2012).

FINE-GRAINED CLASSIFICATION FOR BACTERIAL AND VIRAL COMMUNITIES

On the other side of the spectrum, the problem of fine-grained characterization of datasets produced by targeted sequencing has seen great progress in recent years. Contrary to the analysis of non-selected and therefore more complex metagenomic samples, efficient methods have been developed for cases where certain components of microbial communities are experimentally targeted (bacterial or viral). This has been an effective way to circumvent the difficulty of the first-level assignment, albeit without solving it.

For bacterial communities the most efficient solution is to perform a tag survey, where only partial genomic information is used and the sequencing is performed for marker genes, such as 16S rRNA for prokaryotes and 18S rRNA for eukaryotes (fungi).

This simplifies the analysis for two reasons. First, the amount of data remains reasonable (for a high-throughput analysis) and second, known marker genes' taxonomic classification is available through reference taxonomies such as RDP (Cole et al., 2009) or Greengenes (McDonald et al., 2012). Sequence similarity techniques combined with reference taxonomies recapitulate the known distribution of bacterial phyla extremely well (Bazinet and Cummings, 2012). However, this type of analysis has one major pitfall: it does not provide a reliable method to quantify the identified species (Roux et al., 2011).

While this approach is feasible for bacterial populations, it is not applicable for the analysis of viral communities due to the absence of such marker genes (e.g., Edwards and Rohwer, 2005). Virome studies concentrate on the viral part of the environmental sample and isolate viral genomes encapsidated in viral particles that are purified by a combination of filtration and (utlra)centrifugation. This now popular approach drastically reduces the complexity of the community, which makes it possible to assemble longer contigs routinely (10 kb and more), and even complete genomes from low-complexity samples (Coetzee et al., 2010; Minot et al., 2012). However, it does not really solve the problem of first-level assignment but merely sidesteps it: given the purification step, all generated sequences are generally considered "by definition" as viral, unless proven otherwise by homology-based approaches. In addition, this strategy is not without some caveats (see for more details Fancello et al., 2012). For example, the purified particles may contain cellular genome fragments rather than viral genomes, because of the presence of GTA (Lang and Beatty, 2007) or as a consequence of generalized transduction (for a review see Frost et al., 2005). Also, while 0.22 μ filtering avoids contamination by bacterial, archaeal or eukaryotic cells, other DNA-containing elements, such as bacterial vesicles (Biller et al., 2014) may co-purifiy with virions. Such filtering-based purification also excludes the largest viruses and therefore results in an incomplete picture of viral diversity. Moreover, both LA (see Duhaime et al., 2012) and MDA amplifications have their downfalls. For the former, adapter ligation is only possible for dsDNA viruses and hence ssDNA viral genomes are mostly absent in the sample. For the latter, the amplification is preferentially performed for circular ssDNA viruses rather than dsDNA. The effect of the presence of cellular genes on the bioinformatics analysis of viral metagenomic data has been described and some approaches to detect their presence have been proposed (Roux et al., 2014).

Notwithstanding, virome studies have seen large success. In contrast with bacterial communities, alignment-based methods do not seem to be best suited for viral classification. Indeed, as mentioned by Suttle (2007) even for relatively long viral reads the homolog frequency between these reads and protein sequences within the Genbank database is only about 30%. The idea is to avoid the strong sequential constraint imposed by alignment methods on nucleotides' similarity and to capture a global similarity signal based on sequence composition (*k*-mers). Composition-based techniques seem to provide satisfactory results for fine-grained taxonomic classification of filtered viral samples (e.g., Yang et al., 2005; Trifonov and Rabadan, 2010).

WHY IS THE FIRST-LEVEL ASSIGNMENT PROBLEM HARD?

As we observed in the previous sections, methods for first-level and fine-grained assignment of metagenomic samples co-exist, but exhibit drastically different performances. This naturally raises the question of reasons underlying this performance gap. Since the characterization of metagenomic samples can be formulated as a supervised machine-learning task, we propose here to employ data complexity and hardness measures to compare the intrinsic difficulty of classifying metagenomic samples at the first-level with that of fine-grained assignment.

We consider here three classification tasks whose goals are to assign a class label to each instance of a set of sequences. The three tasks we describe vary by the composition of the set of sequences and by the scope of the class labels to assign.

(1) Given a sample of bacterial sequences, to assign each of them to a phylum (e.g., *Proteobacteria*) or to a class (e.g., *Gammaproteobacteria*);
(2) Given a sample of viral sequences, to assign each of them to a group (e.g., dsDNA) or to a family (e.g., *Plasmaviridae*); and
(3) Given a sample of sequences, to assign each of them to a first-level domain (e.g., bacteria, archaea, eukaryota, or virus).

The classification tasks (1) and (2) are fine-grained assignment problems and mimic characterization of targeted metagenomic studies; while task (3) represents a first-level assignment and mimics the analysis of complex, untargeted environmental samples. Since we are interested in the identification of novel species in large metagenomic samples, we adopted the representation of sequences as *k*-mer frequency vectors.

We analyzed these three classification tasks using an instance-level analysis of data complexity. In supervised machine learning, the performance of a classifier is dependent both on the learning algorithm (e.g., SVM or Naïve Bayes) and on the training data. While global metrics recapitulate overall performances of a classifier, they fail to indicate whether moderate performances are a consequence of wrong parameter adjustments, biased resampling for training data or of the intrinsic difficulty of the classification task. However, recent literature on instance misclassifications demonstrates that for a given classification task, some instances are intrinsically hard to classify and that their presence is indicative of the global difficulty (see Smith et al. (2014) for a review). Most studies agree on the hardness of outlier instances or on instances belonging to a minority class, but Smith demonstrated that simple metrics can actually quantify the intrinsic hardness of an instance. One of these metrics is the *k*-Disagreeing Neighbors (kDN), which measures for a given instance the number of *k* nearest neighbors that do not share its class label. Smith demonstrated that the kDN measure is strongly positively correlated with the misclassification of an instance over a wide range of learning algorithms and of training data resampling.

To compare the classification hardness of the three tasks, we generated from a representative subset of sequenced organisms from Genbank (September 2014 download, 25,624 bioprojects, 100% of viruses, archaea and bacteria, 24 eukaryotes with 18 plants) 100 sets of 10,000 randomly chosen contiguous genomic

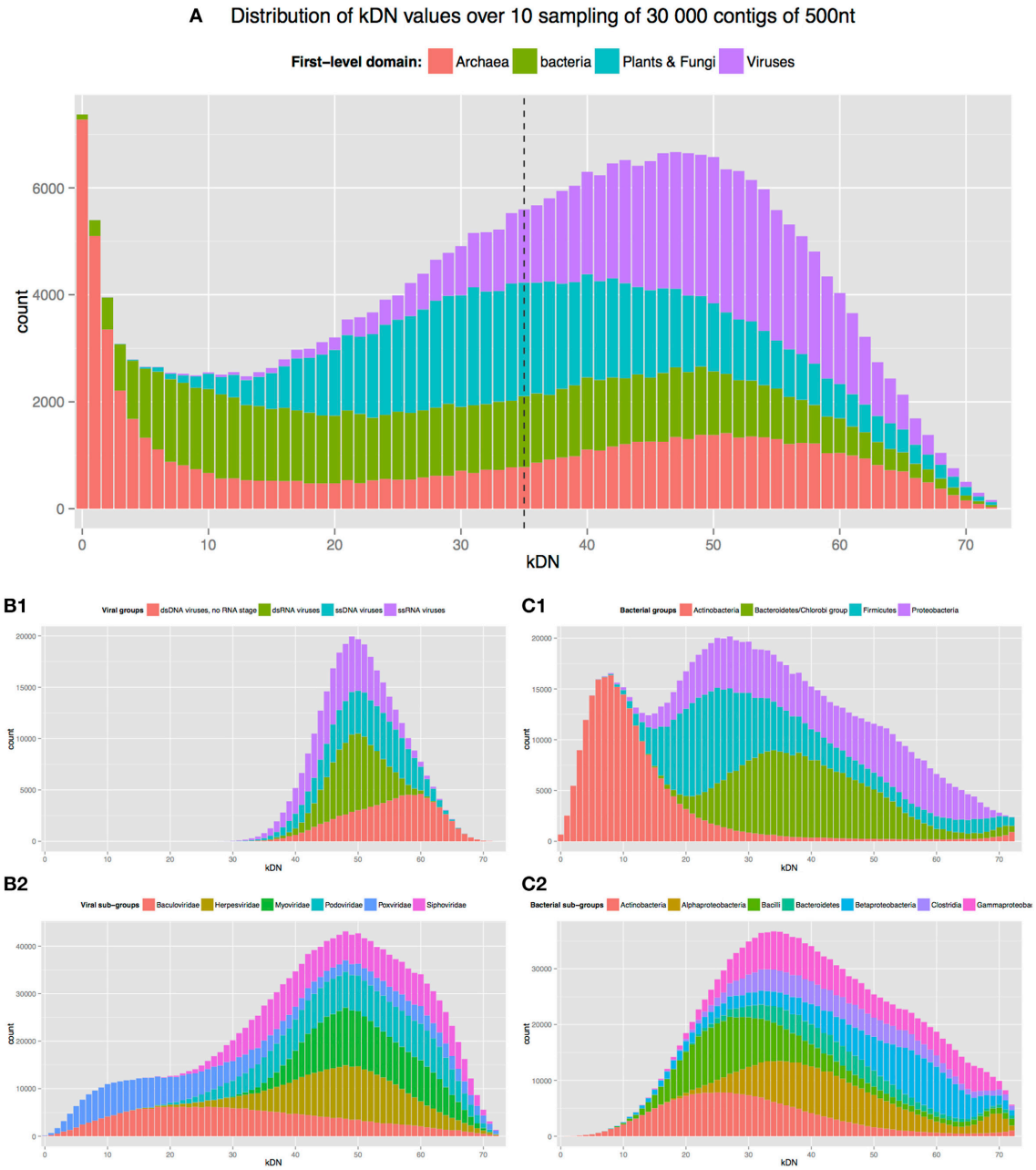


FIGURE 1 | Distribution of kDN by classes for each of three classification tasks. (A) Corresponds to Task 3—assignment of 500 nt contigs to first-level domains; **(B1,B2)** to Task 2—assignment of 500 nt viral contigs to a group or to a family, respectively; **(C1,C2)** to Task 1—assignment of 500 nt bacterial contigs to a phylum or to a class, respectively. Each of the 300,000 randomly selected contigs sampled from different first-level domains were represented as vectors of 3-mer frequencies. Histograms indicate how many contigs (y-axis) per class (colors) have a certain number of neighbors (x-axis) not sharing their own class label, within the closest 73 neighbors. Neighbors are determined w.r.t. euclidean distance in the space of 3-mer frequencies (cf. Section Why is the First-level Assignment Problem Hard? of main text). For example, there are more than 6000 different archaeal contigs (red bar) not having a single non-archeal contigs in their closest 73 neighbors (red bar corresponding to 0 kDN). The dashed line represents the boundary between contigs easy to classify correctly with a majority vote (to the left of the line) and hard to classify (to the right). Only the top 4 most abundant classes are shown for **(B1,C1)**; and 6 for **(B2,C2)**.

fragments of 500 nt average length (corresponding to the average size of metagenomic contigs to simulate an assembly step). For task (1), only bacterial genomes were considered, for task (2) only viral genomes were considered, while for task (3) a balanced composition of viruses, archaea, bacteria and eukaryotes were considered. Each sequence was represented as a 3-mer frequency vector (i.e., the number of time each possible 3-nt sub-sequence appears in the contig) and we defined the distance between two

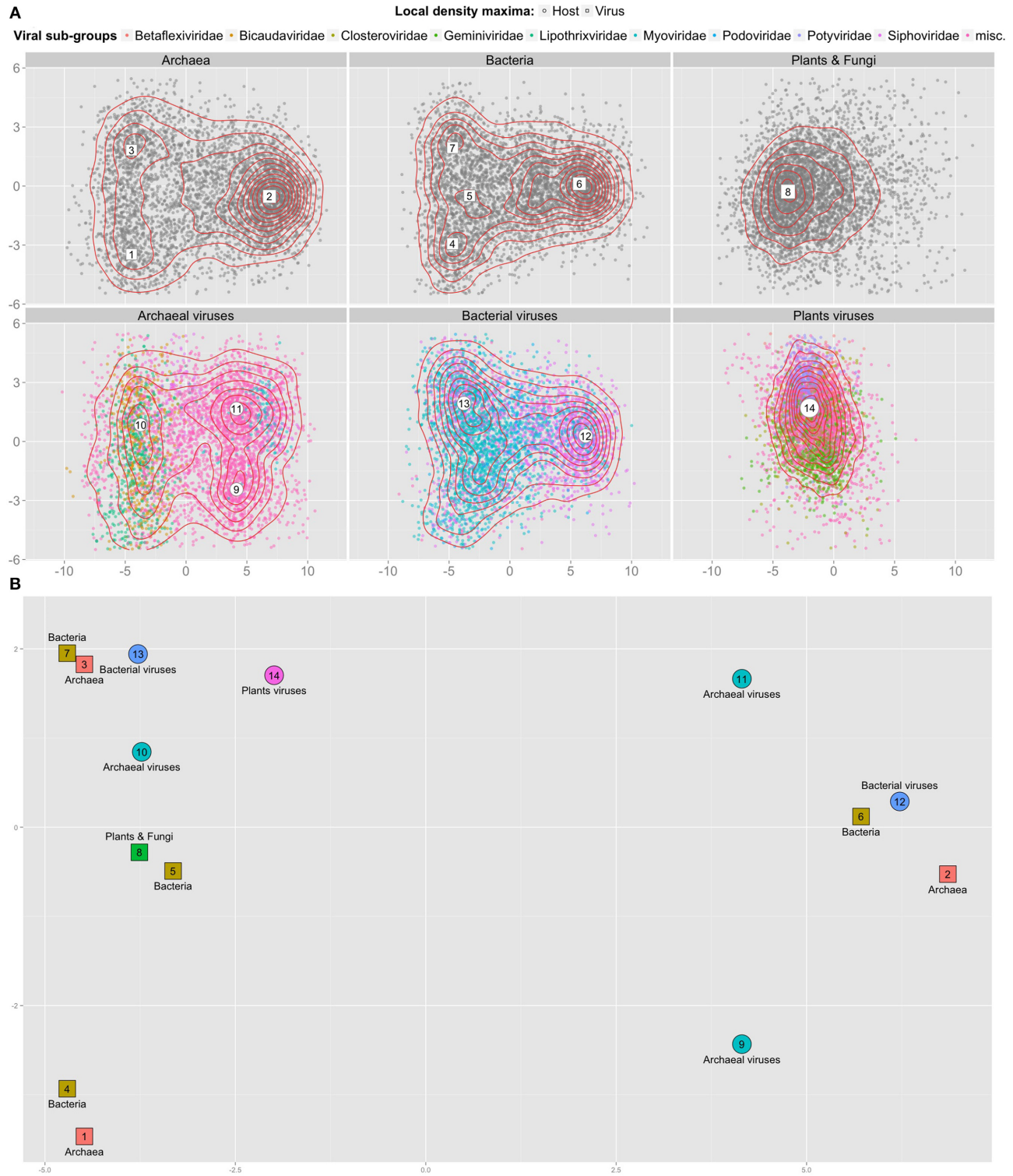


FIGURE 2 | 2D projection of 3-mer frequencies for cellular and viral contigs. (A) Top two dimensions from the PCA reduction of 28,134 contigs (points) of average length 500 nt represented as frequency vectors of 3-mers; sampled equally from genomes originating from 3 top levels cellular domains (top row) and from 3 viral types known to infect them (bottom row). Dimension 1 (x-axis) accounts for 30% of the variance, dimension 2 (y-axis) for 8% of the variance. For each sub-panel, 2 d kernel density estimation is represented using red contour lines and local density maxima are numbered within large white shapes. **(B)** Close up of **(A)** with all local density maxima. The principal components were computed once for the whole set of contigs of all genomes. Position, coordinates and axes from all sub-panels are comparable.

contig as the Euclidean distance between their respective 64 (4^3) dimensional vectors. For each contig, its kDN value is the number of other contigs that do not share its class label among its closest 73 neighbors. The corresponding class hardness is then measured as the median kDN of all the contigs in a given class. We also determined whether an observed median kDN is significantly extreme (low value indicating easy classification, high value corresponding to difficult classes), by estimating the distribution of the median kDN under the null hypothesis of no relation between class labels by random permutations.

We summarize in **Figure 1** the distribution of kDN by class for each of three tasks. The upper panel shows that for the first-level classification task, archaeal and bacterial contigs can be easily assigned to their respective domain, and that this classification is hard for eukaryotic contigs and even harder for viral ones. When the classification task is restricted to bacteria only (panels C1 and C2), fine-grained classification is not hard at both phylum and class levels. For viruses (panels B1 and B2), fine-grained classification to groups (ssDNA, dsRNA etc.) is hard, while assigning a viral sequence to a family level is easier, though less easy than for bacteria. Using a permutation scheme, we established that the observed kDN value is significantly different from the null kDN values for all but the virus fine-grained classification to groups (data not shown).

Consistent with previous work (Mende et al., 2012; Teeling and Glockner, 2012), we have verified that for contigs shorter than 500 nt, distributions are shifted to the right—which corresponds to a harder classification problem (data not shown); conversely, for contigs longer than 500 nt, distributions are shifted to the left, corresponding to an easier classification problem (see Supplementary Figure 1).

It has been previously observed that viral 3-mer signatures are close to that of their hosts (Pride et al., 2006). However, evidence contradicting this observation has also been proposed, for example for large viruses (Mraìzek and Karlin, 2007) and for viruses of monocots and dicots (Adams and Antoniw, 2004). We investigated whether the classification difficulty could be explained by overlapping *k*-mer distributions between different types of hosts and viruses that infect them. To this end we sampled 4689 contigs from each first level cellular groups (archaeal, bacterial, plant and fungal genomes); and of viruses known to infect them. Using principal component analysis (PCA), we projected the 3-mers frequencies vectors of these contigs on 2 dimensions. **Figure 2** shows that viral and cellular contigs are spread uniformly in these 2 dimensions, with the exception of plants viruses that are more compact. Using local density analysis, we observed that contigs of bacterial viruses indeed are close to their hosts (points 12 and 6, 13, and 7), but that they are also as close to archaeal contigs (points 13 and 3). On the other hand, archaeal viruses are not close to their hosts; while plant viruses are closer to bacteria and archaea than to their hosts.

DISCUSSION

Distinguishing viral and cellular sequences in non-targeted environmental studies is a yet unresolved classification problem, especially for unknown viral species. We have shown that the reason why this problem has been eluding a satisfactory solution lies in its intrinsic computational difficulty. The reason for this difficulty lies in the fact that viral sequences *k*-mer distributions overlap with cellular one's almost indiscriminately. This is to be contrasted with the relative ease of the corresponding classification task for archaea and bacteria that certainly underlies the success of bacterial taxonomic assignment studies. The difficulty for viral sequence classification will be alleviated as the public sequence databases become further populated with acquired viral data but this will not provide a sufficient solution to the problem of novel species discovery.

We strongly believe that appropriate choice of computational methodology and further research efforts in this direction are key for the advancement of this field. In the current state of knowledge, we recommend adopting the strategy of contig-level assembly of reads combined with k-mer frequency-based analysis for the identification of viral sequences in metagenomic samples. As for the development of new methods, the promising avenue for the discovery of novel viral sequences seems to be the relaxation of the stringency of long *k*-mer indexing.

ACKNOWLEDGMENTS

This work was supported in part by the SIRIC BRIO (Site de Recherche Intégrée sur le Cancer-Bordeaux Recherche Intégrée Oncologie). The authors would like to thank Patricia Thébault for helpful discussions as well as the referees for their constructive comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00739/abstract

REFERENCES

Adams, M. J., and Antoniw, J. F. (2004). Codon usage bias amongst plant viruses. *Arch. Virol.* 149, 113–135. doi: 10.1007/s00705-003-0186-6

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B., and Allen, J. E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics* 29, 2253–2260. doi: 10.1093/bioinformatics/btt389

Bazinet, A., and Cummings, M. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics* 13:92. doi: 10.1186/1471-2105-13-92

Biller, S. J., Schubotz, F., Roggensack, S. E., Thompson, A. W., Summons, R. E., and Chisholm, S. W. (2014). Bacterial vesicles in marine ecosystems. *Science* 343, 183–186. doi: 10.1126/science.1243457

Blottière, H. M., de Vos, W. M., Ehrlich, S. D., and Doré, J. (2013). Human intestinal metagenomics: state of the art and future. *Curr. Opin. Microbiol.* 16, 232–239. doi: 10.1016/j.mib.2013.06.006

Cénit, M. C., Matzaraki, V., Tigchelaar, E. F., and Zhernakova, A. (2014). Rapidly expanding knowledge on the role of the gut microbiome in health and disease. *Biochim. Biophys. Acta* 1842, 1981–1992. doi: 10.1016/j.bbadis.2014.05.023

Claverie, J. M., Abergel, C., and Ogata, H. (2009). Mimivirus. *Curr. Top. Microbiol. Immunol.* 328, 89–121. doi: 10.1007/978-3-540-68618-7_3

Coetzee, B., Freeborough, M.-J., Maree, H. J., Celton, J.-M., Rees, D. J. G., and Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. *Virology* 400, 157–163. doi: 10.1016/j.virol.2010.01.023

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., et al. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–D145. doi: 10.1093/nar/gkn879

Duhaime, M. B., Deng, L., Poulos, B. T., and Sullivan, M. B. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* 14, 2526–2537. doi: 10.1111/j.1462-2920.2012.02791.x

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

Fancello, L., Raoult, D., and Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology* 434, 162–174. doi: 10.1016/j.virol.2012.09.025

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732. doi: 10.1038/nrmicro1235

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* 195, 194–204. doi: 10.1016/j.jviromet.2013.08.035

King, A. M. Q., Adams, M. J., Carstens, E. B., and Lefkowitz, E. J. (2012). *Virus Taxonomy, Classification and Nomenclature of Viruses*. Amsterdam: Elsevier Academic Press.

Lang, A. S., and Beatty, J. T. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. *Trends Microbiol.* 15, 54–62. doi: 10.1016/j.tim.2006.12.001

Lecuit, M., and Eloit, M. (2014). The human virome: new tools and concepts. *Trends Microbiol.* 21, 510–515. doi: 10.1016/j.tim.2013.07.001

Manor, O., Levy, R., and Borenstein, E. (2014). Mapping the inner workings of the microbiome: genomic- and metagenomic-based study of metabolism and metabolic interactions in the human microbiome. *Cell Metab.* 20, 742–752. doi: 10.1016/j.cmet.2014.07.021

McDonald, D., Price, M., Goodrich, J., Nawrocki, E., DeSantis, T., Probst, A., et al. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* 6, 610–618. doi: 10.1038/ismej.2011.139

Mende, D., Waller, A., Sunagawa, S., Jarvelin, A., Chan, M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* 7:2. doi: 10.1371/journal.pone.0031386

Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2012). Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3962–3966. doi: 10.1073/pnas.1119061109

Mraìzek, J., and Karlin, S. (2007). Distinctive features of large complex virus genomes and proteomes. *Proc. Natl. Acad. Sci. U.S.A.* 104, 5127–5132. doi: 10.1073/pnas.0700429104

Nalbantoglu, O. U., Way, S. F., Hinrichs, S. H., and Sayood, K. (2011). RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics* 12:41. doi: 10.1186/1471-2105-12-41

Ni, J., and Tokuda, G. (2013). Lignocellulose-degrading enzymes from termites and their symbiotic microbiota. *Biotechnol. Adv.* 31, 838–850. doi: 10.1016/j.biotechadv.2013.04.005

Petrosino, J., Highlander, S., Luna, R., Gibbs, R., and Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* 55, 856–866. doi: 10.1373/clinchem.2008.107565

Pride, D. T., Wassenaar, T. M., Ghose, C., and Blaser, M. J. (2006). Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacterio- phages and eukaryotic viruses. *BMC Genomics* 7:8. doi: 10.1186/1471-2164-7-8

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annu. Rev. Genet.* 46, 359–369. doi: 10.1146/annurev-genet-110711-155600

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 1, 289–297. doi: 10.1016/j.coviro.2011.06.004

Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2010). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27, 127–129. doi: 10.1093/bioinformatics/btq619

Roux, S., Enault, F., Bronner, G., and Debroas, D. (2011). Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS Microbiol. Ecol.* 78, 617–628. doi: 10.1111/j.1574-6941.2011.01190.x

Roux, S., Tournayre, J., Mahui, A., Debroas, D., and Enault, F. (2014). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15:76. doi: 10.1186/1471-2105-15-76

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Mach. Learn.* 95, 225–256. doi: 10.1007/s10994-013-5422-z

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

Teeling, H., and Glockner, F. (2012). Current opportunities and challenges in microbial metagenome analysisc—bioinformatic perspective. *Brief. Bioinformatics* 13, 728–742. doi: 10.1093/bib/bbs039

Trifonov, V., and Rabadan, R. (2010). Frequency analysis techniques for identification of viral genetic data. *MBio J.* 1:e00156-10 . doi: 10.1128/mBio.00156-10

Vayssier-Taussat, M., Albina, E., Citti, C., Cosson, J., Jacques, M. A., Lebrun, M. H., et al. (2014). Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics. *Front. Cell. Infect. Microbiol.* 4:29. doi: 10.3389/fcimb.2014.00029

Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Wooley, J., and Yuzhen, Y. (2009). Metagenomics: facts and artifacts, and computational challenges. *J. Comput. Sci. Technol.* 25, 71–81. doi: 10.1007/s11390-010-9306-4

Yang, A., Goldberger, A., and Peng, C.-K. (2005). Genomic classification using an information-based similarity index: application to the sars coronavirus. *J. Comput. Biol.* 12, 1103–1116. doi: 10.1089/cmb.2005.12.1103

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

Received: 10 October 2014; accepted: 05 December 2014; published online: 07 January 2015.

*Citation: Soueidan H, Schmitt L-A, Candresse T and Nikolski M (2015) Finding and identifying the viral needle in the metagenomic haystack: trends and challenges. Front. Microbiol. **5**:739. doi: 10.3389/fmicb.2014.00739*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in
MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 January 2015
doi: 10.3389/fmicb.2014.00790

Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to a *Cicer arietinum* population

Walter Pirovano[1]*, Laura Miozzi[2], Marten Boetzer[1] and Vitantonio Pantaleo[3]*

[1] *Genome Analysis and Technology Department, BaseClear B. V., Leiden, Netherlands*
[2] *Institute for Sustainable Plant Protection of National Research Council, Torino, Italy*
[3] *Institute for Sustainable Plant Protection of National Research Council, Bari Research Unit, Bari, Italy*

Edited by:
Alejandro Reyes, Universidad de los Andes, Colombia

Reviewed by:
Carmen Hernandez, Consejo Superior de Investigaciones Científicas, Spain
Vicente Pallas, Instituto de Biologia Molecular y Celular de Plantas – Polytechnic University of Valencia-Spanish National Research Council, Spain

****Correspondence:***
Walter Pirovano, Genome Analysis and Technology Department, BaseClear B. V., Einsteinweg 5, 2333CC Leiden, Netherlands
e-mail: walter.pirovano@baseclear.nl;
Vitantonio Pantaleo, Institute for Sustainable Plant Protection of National Research Council, Bari Research Unit, Via Amendola 122/d, 70126 Bari, Italy
e-mail: vitantonio.pantaleo@cnr.it

Over the past years deep sequencing experiments have opened novel doors to reconstruct viral populations in a high-throughput and cost-effective manner. Currently a substantial number of studies have been performed which employ next generation sequencing techniques to either analyze known viruses by means of a reference-guided approach or to discover novel viruses using a *de novo*-based strategy. Taking advantage of the well-known *Cymbidium ringspot virus* we have carried out a comparison of different bioinformatics tools to reconstruct the viral genome based on 21–27 nt short (s)RNA sequencing with the aim to identify the most efficient pipeline. The same approach was applied to a population of plants constituting an ancient variety of *Cicer arietinum* with red seeds. Among the discovered viruses, we describe the presence of a *Tobamovirus* referring to the *Tomato mottle mosaic virus* (NC_022230), which was not yet observed on *C. arietinum* nor revealed in Europe and a viroid referring to *Hop stunt viroid* (NC_001351.1) never reported in chickpea. Notably, a reference sequence guided approach appeared the most efficient in such kind of investigation. Instead, the *de novo* assembly reached a non-appreciable coverage although the most prominent viral species could still be identified. Advantages and limitations of viral metagenomics analysis using sRNAs are discussed.

Keywords: bioinformatics, chickpea, ancient varieties, plant viruses, reference sequences, *de novo* assembly

INTRODUCTION

Over the past years deep sequencing experiments have opened novel doors to reconstruct viral populations in a high-throughput and cost-effective manner (Barba et al., 2014; Massart et al., 2014). Currently a substantial number of studies have been performed which employ next generation sequencing (NGS) techniques to either analyze known plant viruses by means of a reference-guided approach or to discover novel plant viruses using a *de novo*-based strategy (Kreuze et al., 2009, 2013; Navarro et al., 2009; Szittya et al., 2010; Wu et al., 2010; Giampetruzzi et al., 2012; Loconsole et al., 2012; De Souza et al., 2013; Candresse et al., 2014; Seguin et al., 2014; Marais et al., 2015). Despite of the significant advances made by sequencing technologies only a few methods have been proposed to specifically analyze viral metagenomes, especially if compared to the number of tools designed for, e.g., bacterial metagenome analysis (Schloss et al., 2009; Huson et al., 2011). At least in part this can be attributed to the fact that most viruses are still undiscovered; it has been suggested that at present less than 1% of the extent of viral diversity has been explored (Mokili et al., 2012). Moreover, viral communities tend to be phylogenetically distant from each other and therefore virus discovery and reconstruction heavily relies on *de novo* approaches. Another hurdle resides in the fact that viral populations are highly heterozygous which is mainly due to the low fidelity of the viral polymerase. This leads inevitably to a high mutation rate and consequently high variation even within the same populations that comprise a viral quasispecies (Domingo et al., 2012). Assembly tools of short sequence reads such as De Bruijn graph-based methods (Zerbino and Birney, 2008) are in principle designed for linear assembly of less diverse haploid and diploid genomes. As a result the assembly of viral (meta)genomes often leads to a substantial amount of contigs with generally a very short average length. Thus subsequent amplification of the resulting fragments using traditional methods (such as PCR and Sanger sequencing) is often essential to extend the draft assembly.

Also it should be mentioned that the chance of properly reconstructing one or more viral taxonomies heavily depends on the quantity of viral genomes present in the input sample. Given that viruses cannot easily be isolated, generally a high sequencing coverage is necessary to pick up all relevant viral genomic material within a plant sample. Alternatively, virus enrichment is needed (Roossinck, 2012). In other words, projects that aim to characterize viral (meta)genomes in plants can become very costly as these are mostly based on sequencing total DNA or RNA libraries that contain only a small fraction of viral material.

The silencing-based antiviral plant response may help somehow in this deal; it implies the recognition of double-stranded

(ds) or ds-like RNAs of viral origin by members of plant Dicers (DCLs; Aliyari and Ding, 2009). The recognized viral RNAs are then processed by DCLs into viral small interfering RNAs (v-siRNAs; reviewed by Ding and Voinnet, 2007 and Ruiz-Ferrer and Voinnet, 2009). Two distinct classes of v-siRNAs have been identified: primary v-siRNAs, which result from the DCL mediated cleavage of an initial trigger RNA, and secondary v-siRNAs, which require a plant RNA-directed RNA polymerase (RDR) for their biogenesis (Wassenegger and Krczal, 2006; Donaire et al., 2008; Ruiz-Ferrer and Voinnet, 2009; Vaistij and Jones, 2009; Garcia-Ruiz et al., 2010; Wang et al., 2010). The amplification and high level of v-siRNAs accumulation in many but not all virus infections depends on the combined activity of the host-encoded RDRs such as RDR1, RDR2, and RDR6 with other factors such as the RNA helicase SDE3. The amplification mechanism may result in production of secondary amplified v-siRNAs also in case of weakly induced silencing (i.e., low accumulation of viral RNAs; Garcia et al., 2012).

v-siRNAs can also be successfully used to cover known viral genomes by aligning reads to the reference sequences (ref_seq), thus providing a simple method for detection of known viruses and viroids and their variants (Navarro et al., 2009; Pantaleo et al., 2010). In addition, Kreuze et al. (2009) have used at first sRNA libraries for *de novo* reconstruction of the complete genome of a known plant RNA virus from multiple contigs of v-siRNAs. Moreover v-siRNAs can be used for non-homologous discovery of novel plant infectious entities (Wu et al., 2012). The deepness and the low level of bias of sRNAs are key factors for the success of either reference alignment and *de novo* assembly based approaches. Seguin et al. (2014) have demonstrated that is possible to reconstruct the entire genomic master sequence of DNA and RNA viruses from both model and crop plants using v-siRNA libraries when sequencing approximately 20 million deep sRNA libraries. Other research groups have spent efforts to demonstrate that bias in cloning procedures may hide some of the sRNAs and therefore they have studied and developed alternative strategies to reduce such bias (Sorefan et al., 2012).

In the present paper we analyze a specific sRNA library from leaves sampled within plants constituting a *Cicer arietinum* ancient variety (Red of Ruvo, Apulia-Italy) and from leaves of *Nicotiana benthamiana* plants infected with the *Cymbidium ringspot virus* (*CymRSV*) in a ratio of approximately 1000 to 1. The presence of v-siRNAs from *CymRSV* allows us to compare different bioinformatics tools developed for reference-guided or *de novo* assembly based approaches. Protocols were also applied to the viral metagenome of *C. arietinum*. We find that a reference-guided approach is very successful in the reconstruction of the most abundant viruses. Instead *de novo* approaches clearly suffer from the heterogeneity within viral populations. Among the discovered viruses, we describe the presence of a *Tobamovirus* referring to the *Tomato mottle mosaic virus (ToMMV;* NC_022230), which was not yet observed on *C. arietinum* and also not yet revealed in Europe, and one viroid referring to *Hop stunt viroid (HSVd;* NC_001351.1) never reported in chickpea. Accordingly, we discuss our findings and provide suggestions that aim to discover plant–viruses using a cost-effective approach based on sRNA sequencing.

MATERIALS AND METHODS

PLANT MATERIALS, VIRUS, RNA EXTRACTION, AND SMALL RNA SEQUENCING

The use of wild type *N. benthamiana* plants and infection with *CymRSV in vitro* transcripts was previously described (Pantaleo et al., 2007; Pantaleo and Burgyan, 2008). The plant growth chamber was set with 10 h in light and 14 h in dark at 22°C. Seed population constituting an ancient variety of *C. arietinum* named "Red Chickpea of Cassano delle Murgie" accession "Red of Ruvo" (in collection at Mediterranean Germplasm Database, http://ibbr.cnr.it/ibbr/resources/mediterranean-germplasm-database) was grown in an open air collection field. Plant leaf material representing the entire population (i.e., one leaf per plant covering the 30% of the plants) was collected at flowering stage and bulked. Total RNA was extracted from plant tissues using Tri-Reagent (SIGMA) following manual instructions. Low molecular weight RNA was enriched as previously described (Johansen and Carrington, 2001) and mixed in a ratio of 1000 (chickpea) to 1 (*N. benthamiana*) in amount. Subsequently, libraries of sRNAs were produced using a TruSeq Small RNA Sample Kit (Illumina) and sequenced with standard sequencing oligos on the Illumina HiSeq 2500 platform. Short sequence reads were generated using bcl2fastq software (v 1.8.3). The dataset has been deposited in GEO Omnibus under the entry code GSE63378.

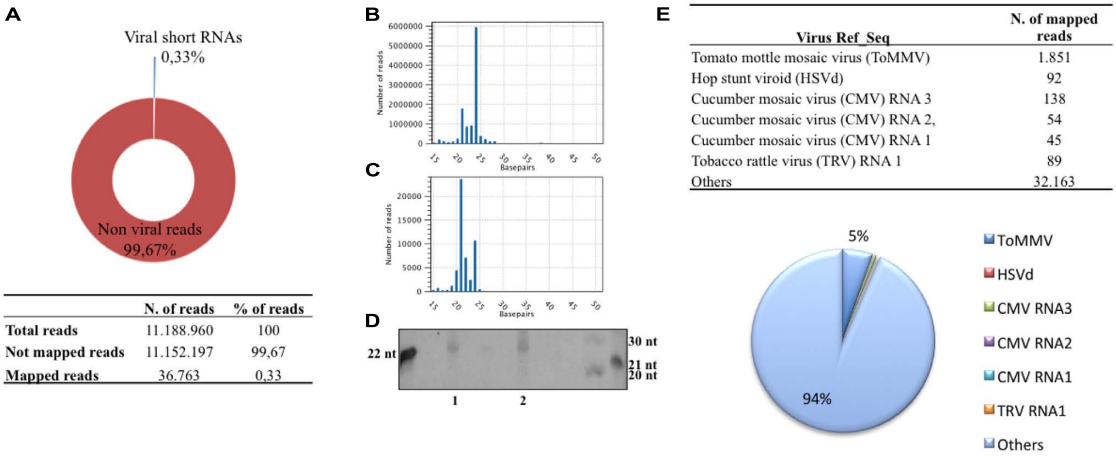
BIOINFORMATICS

Small RNA adapters were removed from the Illumina sequence reads using the "Trim sequences" option of the CLC Genomics Workbench (v 6.0.4). For the ref-seq based approach, the resulting sub-reads were aligned against the *CymRSV* (NCBI accession code NC_003532) and *ToMMV* (NCBI accession code NC_022230) reference genomes using the "Map reads to reference" option of the CLC Genomics Workbench (v 6.0.4). For the *de novo* based approach assemblies were generated, respectively, with Velvet version 1.2.10 (Zerbino and Birney, 2008), Oases version 0.2.08 (Schulz et al., 2012), and MetaVelvet (Namiki et al., 2012). Alignment of the assembled contigs against the *CymRSV* and *ToMMV* reference genomes was performed using Burrows-Wheeler Aligner (BWA) v. 0.7.7 (Li and Durbin, 2009). From the alignment consensus sequences were generated using SAMtools version 0.1.19 (Li et al., 2009). SNP detection was performed with Nucmer which is part of the MUMmer analysis package (version 3.22; Kurtz et al., 2004). Graphical alignment visualization were generated using the Integrative Genomic Viewer (IGV; Robinson et al., 2011). All software was used with default settings unless otherwise specified in text and figures.

RESULTS

SHORT RNA DATASET

sRNAs specifically of 20–27 nucleotides with 5′-phosphate and 3′-OH (likely to be DCL products) were isolated from *C. arietinum* "Red of Ruvo" and from a *CymRSV*-infected *N. benthamiana* and further identified by high-throughput Illumina sequencing. The library yielded in total more than 11 million reads (table in **Figure 1A**) with a minimum and maximum length of 16 and 27 nt (**Figure 1B**). A consistent fraction of these (approximately 6 million) were 24 nt in length (**Figure 1B**) and this is in line with



	N. of reads	% of reads
Total reads	11.188.960	100
Not mapped reads	11.152.197	99,67
Mapped reads	36.763	0,33

Virus Ref_Seq	N. of mapped reads
Tomato mottle mosaic virus (ToMMV)	1.851
Hop stunt viroid (HSVd)	92
Cucumber mosaic virus (CMV) RNA 3	138
Cucumber mosaic virus (CMV) RNA 2,	54
Cucumber mosaic virus (CMV) RNA 1	45
Tobacco rattle virus (TRV) RNA 1	89
Others	32.163

FIGURE 1 | Statistics of the short RNA library from *Cicer arietinum* ancient variety "Red of Ruvo." Pie graph and table resuming the dimension of the library and the fraction of sRNA reads mapping with plant virus ref_seq dataset **(A)**, distribution of sRNA length of non-viral reads **(B)** and viral reads **(C),** Ethidium bromide staining of sRNA from *C. arietinum* (1), not infected *Nicotiana benthamiana* (2; **D**), table and pie graph resuming viral reference sequence (ref_seq) represented in the sRNA library **(E)**.

observations in ethidium bromide staining of the polyacrylamide isolation gel (**Figure 1D**, lane 1). The abundance of 24 nt sRNAs found in chickpea also agrees with previous studies showing that in plants, except for a few species, the 24 nt sRNAs are more abundant than the 21 nt class (Rajagopalan et al., 2006; Moxon et al., 2008; Pantaleo et al., 2010). Accordingly, **Figure 1D** shows that the sRNAs from *C. arietinum* (lane 1) and not infected *N. benthamiana* (lane 2) equally migrate as they are of the same size.

Those sRNAs flanked by the 3′ and 5′ TrueSeq Illumina adapters were compared with a plant virus reference dataset (ftp://ftp.ncbi.nih.gov/refseq/release/viral/), which is defined by collection of 1.677 unique plant virus master sequences. In total 36.763 reads could be mapped to the plant virus reference dataset (excluding those from *CymRSV*), thus the v-siRNAs constituted only 0,33% of the entire library (**Figure 1A**). More than half of the v-siRNAs were of length 21 nt (ca. 23.000), whereas those of length 22 and 24 nt were less abundant (of ca. 7.000 and 11.000, respectively; **Figure 1C**). This distribution recapitulates what was previously observed in plant virus infections, particularly in those infected with RNA viruses, and it mirrors the plant DCLs activity involved in RNA-silencing-based antiviral activity (reviewed by Shimura and Pantaleo, 2011). The most represented viral genomes by viral reads comprise the *Tobamovirus ToMMV* (i.e., 1.851 reads), the *Hop stunt viroid (HSVd*; i.e., 92 reads), the *Cucumber mosaic virus (CMV*; i.e., 45, 54, and 138 reads map against CMV RNA1, 2 and 3, respectively) and the RNA 1 of the *Tobacco rattle virus (TRV*; i.e., 98 reads; Table in **Figure 1E**). The above mentioned viral reads all together represented about 4.5% of the entire population of siRNAs that map against the plant virus dataset, indeed most viral reads are scattered in exiguous number across viral ref_seq (i.e., less than 10 unique reads per ref_seq). Moreover, these putative viral siRNAs align to unrelated viruses (i.e., belonging to different viral families) thus not suggesting the need for further investigations.

REFERENCE SEQUENCE-GUIDED ASSEMBLY

As mentioned above and detailed in the "Materials and Methods" section, the sRNA library under analysis included a small fraction of siRNAs from *CymRSV*-infected *N. benthamiana*. Thus, we have first reconstructed the *CymRSV* genome through alignment of the sRNA reads against its ref_seq (NCBI accession code NC_003532). The alignment statistics are shown in **Table 1**. A total of 364.590 sRNAs reads, with an average length of 21 nt, mapped onto the 4.733 nt long ref_seq. Each nucleotide of *CymRSV* was covered by sRNA reads 77,03 times on average and all together the reads were able to reconstruct 99% of the entire genome (the final consensus sequence comprises 4.698 of the original 4.733 nt). Subsequent variant calling revealed the presence of 13 SNPs; such degree of variability between the consensus sequence and the ref_seq is in agreement with previous findings for *Tombusvirus* variability at 3 days after inoculation of an *in vitro* transcript (Russo et al., 1994).

The same approach was used for the reconstruction of *ToMMV* (NCBI accession code NC_022230). This virus was the best represented by viral reads population in chickpea (**Figure 1E**). Individual alignment of all reads against exclusively NC_022230 shows that a total of 1.909 sRNAs (with an average length of 21,55 nt) could be mapped (**Table 2**). Given the reference length

Table 1 | sRNA alignment statistics against the *Cymbidium ringspot virus (CymRSV)* reference sequence (ref_seq) NC_003532.

Length of the reference CymRSV sequence NC_003532	4.733
Number of mapped reads	364.590
Average length	21,08
Average coverage	77,03
Number of nucleotides in consensus sequence	4.698
Fraction of reference covered	0,99
Number of SNPs with NC_003532	13

Table 2 | sRNA alignment statistics against the *Tomato mottle mosaic virus (ToMMV)* ref_seq NC_022230.

Length of the reference ToMMV sequence NC_022230	6.398
Number of mapped reads	1.909
Average length	21,55
Average coverage	6,4
Number of nucleotides in consensus sequence	5.582
Fraction of reference covered	0,87
Number of SNPs with NC_022230	39

of 6.398 nt, each *ToMMV* nucleotide was represented 6,4 times on average; in total 87% of the entire genome was covered at least one time (the consensus sequence covered 5.582 out of 6.398 nt). Finally, the variant calling analysis revealed the presence of 39 SNPs. Notably, the number of SNPs found is sensibly higher than those found in the model system *CymRSV*. This is particularly interesting if we consider the incidence of SNPs in relation to the total *ToMMV* v-siRNAs (i.e., 1.909) versus those of *CymRSV* (364.590). Nonetheless, the large variability encountered in the present metagenomics investigations on field-cultivated plants is fully in line with previous reports for other non-*in vitro* plant/virus systems (Seguin et al., 2014).

The graphic distribution of mapped reads against *CymRSV* and *ToMMV* is reported in (**Figures 2A,B** respectively). The graphic representation shows that *ToMMV* is almost entirely covered by v-siRNAs in a manner that is at least visually similar to that of *CymRSV*, thus reproducing a high genome coverage of 99% (*CymRSV*) and 87% (*ToMMV*) already indicated in **Tables 1** and **2**. Also a viroid referring to *HSVd* (NC_001351.1) was almost entirely reconstructed with only 92 reads when applying

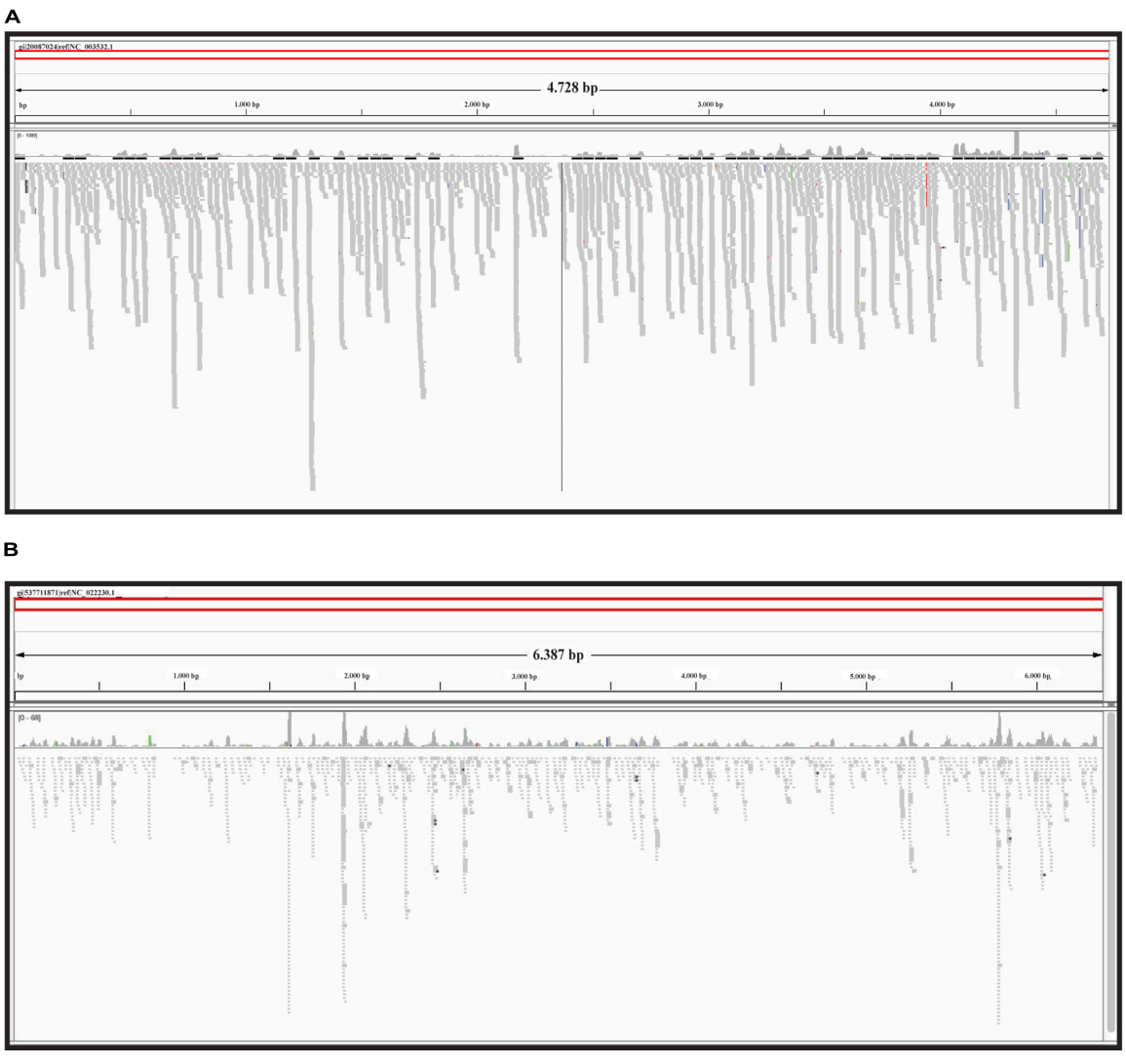


FIGURE 2 | Reference sequence based approach. sRNAs alignment to the *CymRSV* ref_seq NCBI# NC_003532 **(A)** and to *ToMMV* ref_seq NCBI# NC_022230 **(B)**.

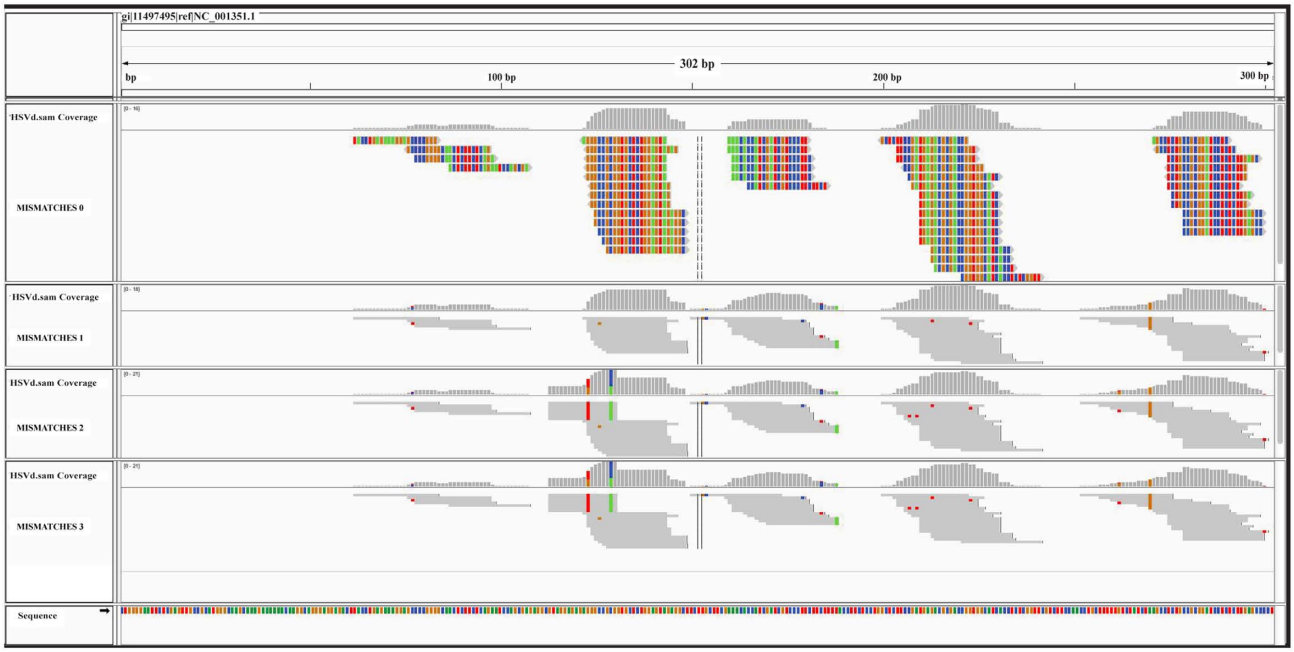


FIGURE 3 | Graphical overview of the ref_seq-based approach. sRNA alignments against the HSVd ref_seq NCBI# NC_001351 are displayed. The four consecutive panels correspond to 0, 1, 2, and 3 mismatches used in BWA alignment. All nucleotides are colored in the expanded graphic representation (mismatches = 0): C in blue, G in brown, A in a green, and T in red. Mismatches are colored in not expanded representations.

the same settings as for *CymRSV* and *ToMMV* (**Figure 3**). The shortness (302 bases) of the viroid allowed us to check whether a better coverage could be obtained by introducing mismatches in BWA alignment protocol. Indeed, some gaps were covered when using two mismatches (**Figure 3**) and at three mismatches no further improvement was obtained. Still, the 5′ part of the viroid (position 1–60 of the released HSVd ref_seq NC_001351), upstream the central conserved domain (CCD; Keese and Symons, 1985; Visvader and Symons, 1985) could not be covered by introducing single variants in the alignment with the ref_seq (see Discussion).

DE NOVO BASED APPROACHES AND v-siRNAs ASSEMBLY

Given that at present ca. 900 species of plant viruses have been determined (Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, 2012) in most cases no good reference (or master consensus genome) is available. In this scenario *de novo* assembly of viral genomes should be considered

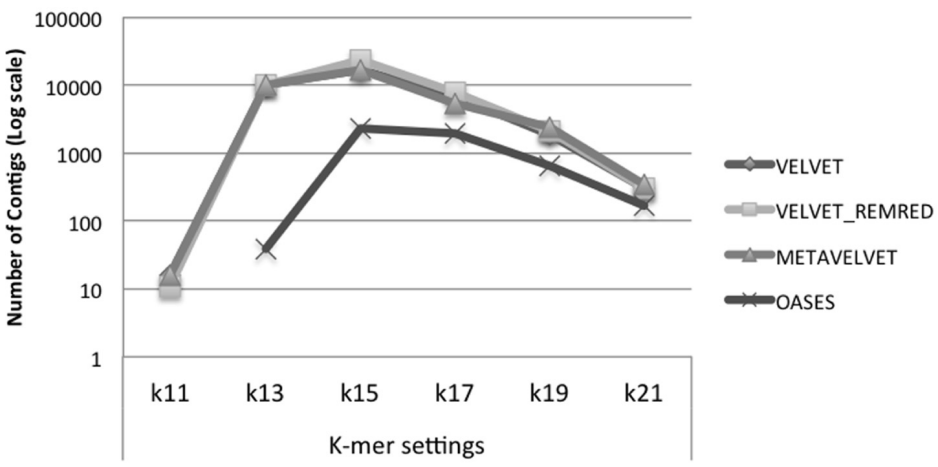


FIGURE 4 | Number of contigs (log scale) obtained using Velvet, Metavelvet, and Oases short (s)RNA assembly tools with different *k*-mer settings. REMRED = Remove redundant.

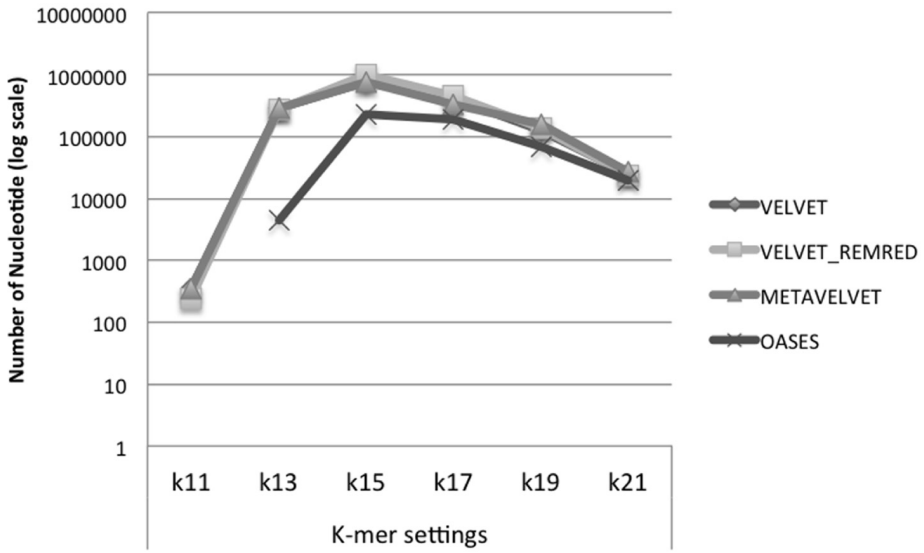


FIGURE 5 | Number of bases (log scale) in the assembled contigs obtained using Velvet, Metavelvet, and Oases short RNA assembly tools with different *k*-mer settings. REMRED = Remove redundant.

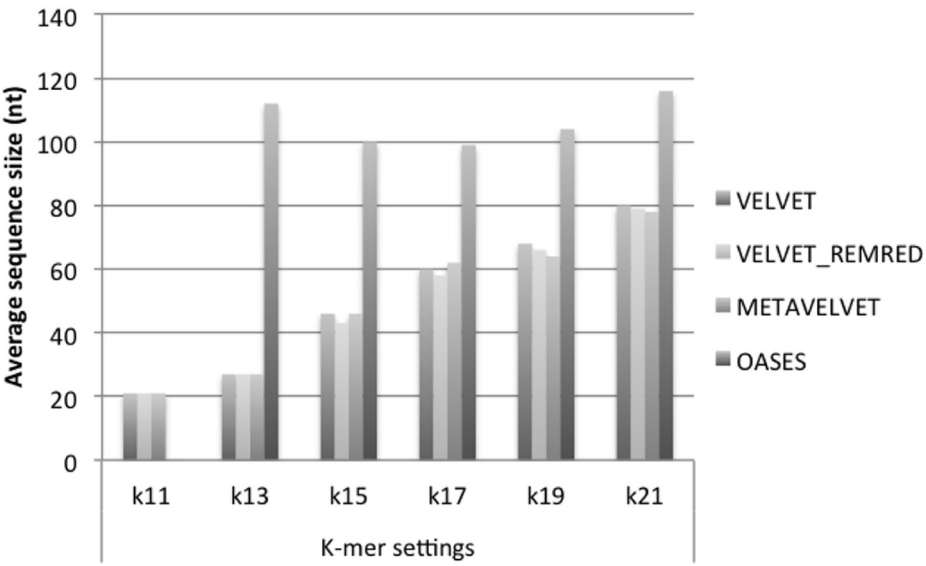


FIGURE 6 | Average size of contigs obtained using Velvet, Metavelvet, and Oases short (s)RNA assembly tools with different *k-mer* settings. REMRED = Remove redundant.

as a valid alternative, thus allowing the creation of a consensus sequence set that best represents the underlying viral population with a non-homology approach. These consensus sequences can serve as a proper basis for reference alignment and variant calling as described above. At present De Bruijn graph-based algorithms (reviewed by MacLean et al., 2009) are the methods of choice to assemble a set of NGS reads. In brief the algorithm divides the NGS reads into short sub-reads (so-called *k*-mers) and subsequently it searches the ideal assembly path in the graph through overlap between the *k*-mers. Thus, the algorithm is optimized for a fast merging of millions of short NGS reads into (large) genomic fragments. In fact, De Bruijn graph based methods such as Velvet (Zerbino and Birney, 2008) and SOAP *de novo* (Li et al., 2010) are widely employed for the genomic assembly of prokaryotes and eukaryotes. The algorithm, however, appears to be less suited for the assembly of fragments with unbalanced coverage distributions such as generated in RNA-Seq and metagenomic libraries. In the latter case chromosomes of different microbes are present in a metagenomics sample proportional to their relative abundance. As such the relative frequencies of short reads

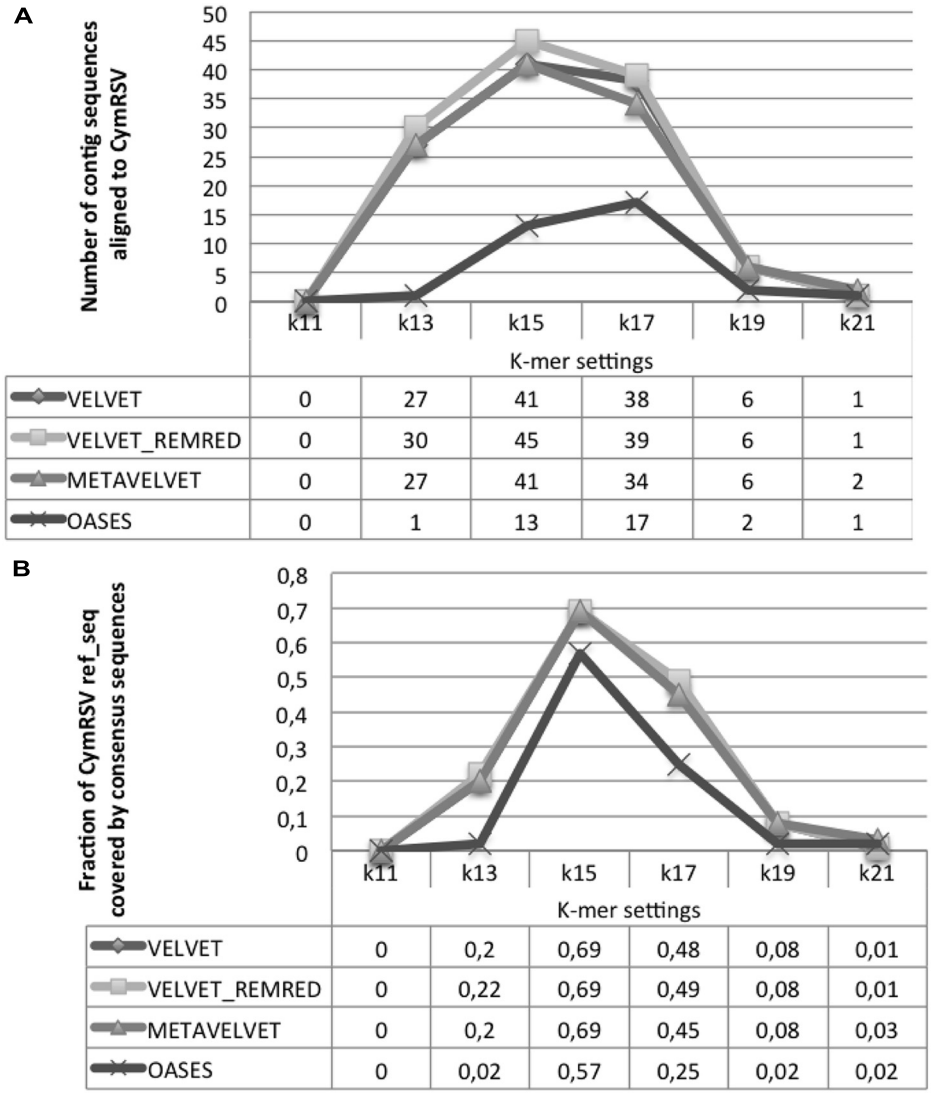


FIGURE 7 | Alignment statistics of contig sequences constructed using Velvet, Metavelvet, and Oases short RNA assembly tools with different *k*-mer settings. The graph and table display the number of contig sequence aligning with *CymRSV* ref_seq **(A)** and fraction of CymRSV ref_seq covered by the consensus sequence **(B)**. REMRED = Remove redundant.

covering the various nodes in the De Bruijn graph differ with respect to a standard linear genome assembly. To overcome these problems specific tools are developed for the assembly of transcriptomes, e.g., Trinity (Grabherr et al., 2011) and Oases (Schulz et al., 2012) and metagenomes [e.g., and MetaVelvet (Namiki et al., 2012) and Ray Meta (Boisvert et al., 2012)]. Potentially these methods could also be of good use for assemblies of viral metagenomes. We here apply the three well-used bioinformatics assembly tools (Velvet, Metavelvet, and Oases) and compare their relative ability to reconstruct the viral metagenome. Also we developed an in-house modification of the standard Velvet protocol where, prior to the assembly, duplicate reads are removed (REDREM) thus taking into account issues related to unbalanced genome coverage (see Discussion). All strategies were evaluated at different *k*-mer settings.

We observe that Velvet and Metavelvet constructed the largest number of consensus sequences at all *k*-mer setting (hereafter "*k*") used. Surprisingly, for all tools the maximum number of consensus sequences was obtained at setting $k = 15$ (**Figure 4**; Supplementary Data 1A). More specifically, Velvet is able to provide a higher number of consensus sequences at $k = 15$ when using a non-redundant sRNA dataset, i.e., 16.604 sequences using Velvet versus 23.251 using Velvet REDREM (Supplementary Data 1A,B, respectively). Accordingly, the total number of assembled nucleotides was higher when using Velvet and Metavelvet compared to Oases (**Figure 5**; Supplementary Data 1).

On the other hand, when comparing the average size of the contig sequences obtained by the different methods at different *k*-mers's, Oases appears to be the best method. Indeed, except in the case of $k = 11$, Oases appears to provide the longest consensus

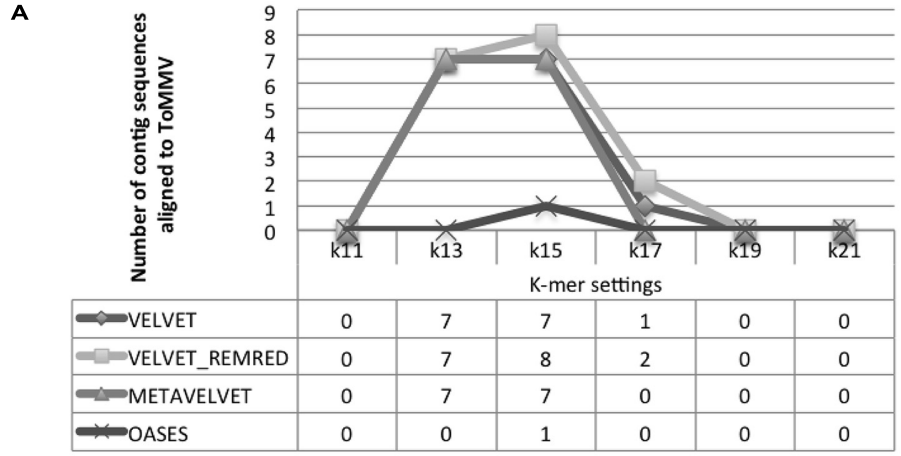


FIGURE 8 | Alignment statistics of contig sequences constructed using Velvet, Metavelvet, and Oases short RNA assembly tools with different *k*-mer settings. The graph and table display the number of contig sequences aligning with ToMMV ref_seq **(A)** and the fraction of ToMMV ref_seq covered by the consensus sequence **(B)**. REMRED = Remove redundant.

sequences on average for all settings (**Figure 6**). Regarding the consensus sequences generated with Velvet and Metavelvet, the average length increased from $k = 11$ (the lowest) to $k = 21$ (the highest; **Figure 6**). The longest consensus sequence was obtained by Oases at $k = 17$ (i.e., 919 nt in length, Supplementary Data 1). Other tools obtained their maximum length (between 400 to 600 nt) at *k*-values ranging from 15 to 21 (Supplementary Data 1).

DE NOVO ASSEMBLY OF CymRSV AND ToMMV

All contig sequences obtained by different tools and settings were aligned against the *CymRSV* ref_seq. First Velvet REMRED and second Metavelvet and Velvet showed to be the most efficient tools by generating, respectively, 45 and 41 consensus sequences at $k = 15$ (**Figure 7A**). For these approaches an increase or a decrease of *k* values resulted in a sensible decrease of contig sequences aligning with the *CymRSV* ref_seq: e.g., in the case of Velvet REMRED the use $k = 13$ or $k = 17$ reduces the number of contigs to 30 and 39, respectively, whereas for $k = 15$ in total 45 contigs were assembled (**Figure 7A**). Moreover, when applying $k = 15$ to both Velvet methods and MetaVelvet the coverage of the *CymRSV* genome was the highest, i.e., 0,69% (3.247 nt out of 4.733 nt of the *CymRSV* genome). Again a setting of $k = 13$ or $k = 17$ sensibly reduces the efficiency of the method (**Figure 7B**).

Oases reached a slightly lower coverage level (57%) at $k = 15$ (**Figure 7B**) although the average length of the 13 consensus sequences was significantly higher than the other three methods (**Figure 7A**).

Surprisingly, all assembly tools evaluated at $k = 15$ detected a similar number of SNPs (14 in case of the Velvet methods, 12 in case of Oases; see Supplementary Data 2), which is comparable to the number of SNPs detected with the reference-guided assembly (i.e., 13, **Table 1**). As previously underlined, the method of *CymRSV* inoculation and the timing of sampling may impede a further increase of variability within the viral genome (Russo et al., 1994).

FIGURE 9 | Alignments of contig sequences obtained using Velvet REMRED, at different *k*-mer settings. Contig distribution graph for CymRSV ref_seq **(A)** and ToMMV ref_seq **(B)** REMRED = Remove redundant.

The alignment of contigs to the *ToMMV* reference genome confirms the findings obtained for *CymRSV*; in **Figure 8A** we show that the $k = 15$ value remains the best setting to obtain highest number of contigs and also the highest coverage (i.e., 0,07, **Figure 8B**). Importantly, contigs obtained at a k value of 15 do not cover the exact same genome segments compared to those obtained with other k-mers (**Figure 9B**). This also holds for *CymRSV* analysis (**Figure 9A**). Thus, assemblies generated at different k values may complement each other to help the reconstruction of more complete viral genomes and increase the coverage as later discussed.

Obviously, a coverage value of 0,07 cannot be considered an acceptable coverage and shows the limitations of such approach for discovering novel viral entities. In **Figure 10** we compare contigs which are *de novo* assembled using different bioinformatics tools at a $k = 15$ settings in the cases of *CymRSV* and *ToMMV* (**Figures 10A,B** respectively). Here we graphically confirm the findings revealed in **Figure 7** and in Supplementary data 2 and 3. Indeed, the genome coverage of *ToMMV* is significantly lower compared to that of *CymRSV* (**Figures 7A,B**) and this is at least partly due to the abundance of viral-deriving siRNAs in the library (i.e., 364.590 vs. 1.909; **Tables 1** and **2**).

DISCUSSION

Viral metagenomics surveys in plants have estimated that only small fractions of virus species are known. Stobbe and Roossinck (2014) have recently proposed the following classification of viruses found by metagenomics: (i) Known–known: virus species or isolates that are already known to be in the environment, (ii) Unknown–known: new virus species or isolates of a known family, or known viruses that have not been found previously in the surveyed environment, and (iii) Unknown–unknowns: viruses that are completely novel and share little to no sequence similarity with other known viruses. In this study we use 10 million sRNAs reads

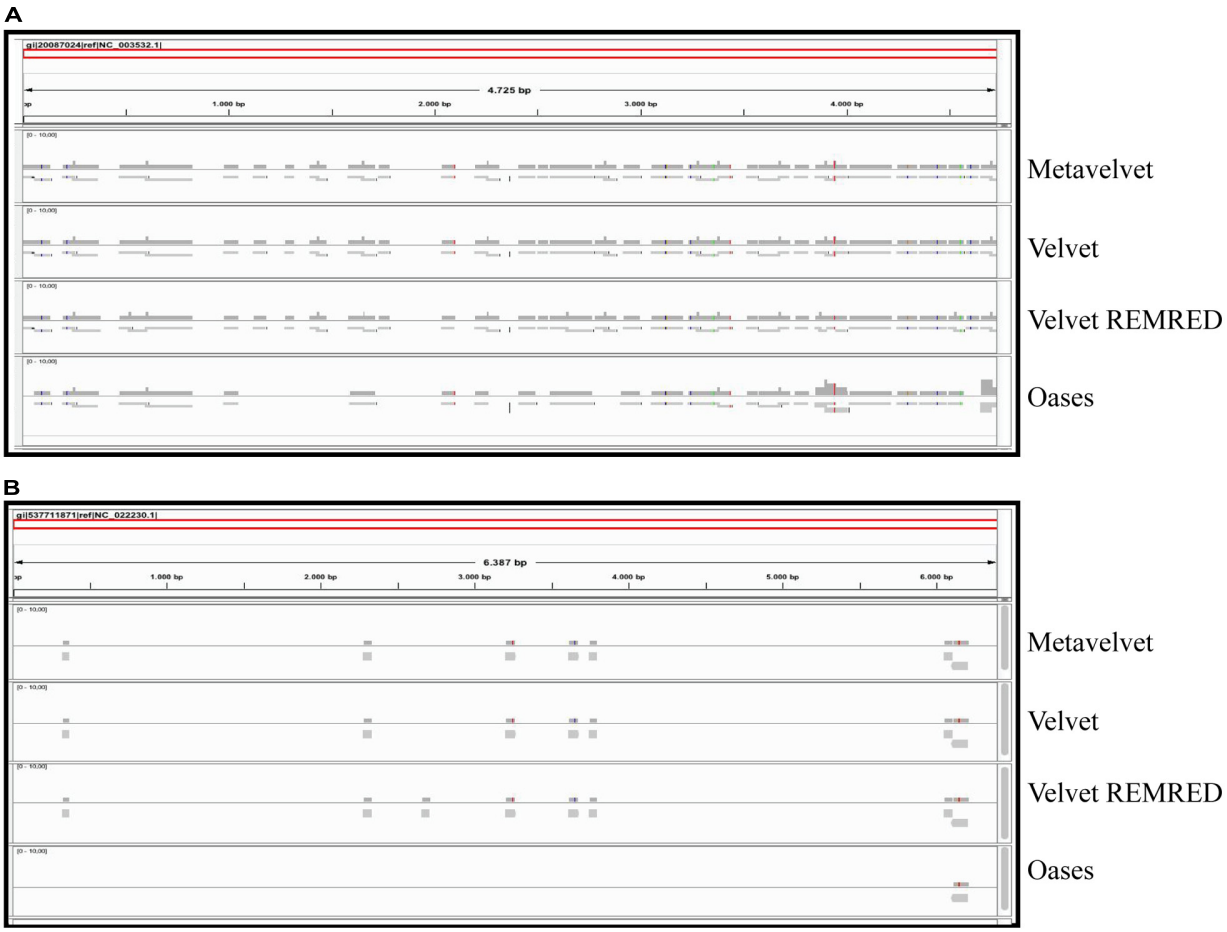


FIGURE 10 | Alignments of contig sequences obtained using Velvet, Metavelvet, and Oases short RNA assembly tools at $k = 15$. Contig distribution graph for CymRSV ref_seq **(A)** and ToMMV ref_seq **(B)** REMRED = Remove redundant are displayed.

library (**Figure 1**) containing CymRSV v-siRNAs. The high coverage level of the virus is indicated by the fact that each nucleotide of the viral genomic ref_seq is represented on average 77 times in sRNAs. Subsequent variant calling showed the presence of a discrete number of SNPs (**Table 1**). Our data describe a typical example of metagenomics analysis of a "known–known" virus in a model-permissive system (i.e., *N. benthamiana*). In this case the ref_seq guided approach is able to reconstruct 99% of the genome whereas the *de novo* based approach is able to cover 69% of the genome. The gap between the two approaches is likely due to the heterogeneity within viral populations, the low and unbalanced genome coverage, and the rather short length of the siRNAs (around 21 bp on average). It is expected that the assembly would at least to some extent benefit from an increased sequencing depth (i.e., by generating 20 million sRNA reads instead of 10 million) although this would of course lead to additional costs for sequencing and data managing (Seguin et al., 2014). Also it should be mentioned that the genome coverage of the *de novo* assembly was calculated based on a reference alignment tool using strict parameters to allow only a few mutations. It is to be expected that more permissive alignment strategies such as BLAST (Altschul et al., 1990) can detect more homologous regions between the assembly and the reference, thus allowing the reconstruction of a more complete genome. However, there is a risk that the allowance of a higher number of mismatches will contemporarily lead to the inclusion of erroneously assembled regions. Further investigations and quality assessment is needed to address this issue.

In the case of short infectious entities such as HSVd, a more permissive alignment strategy obtained by introducing more mismatches in the alignment settings (i.e., 1, 2, or 3), may better be able to cover small gaps (**Figure 3**). However, the approach still leaves un-resolved gaps that could be associated with specific variants in the non-conserved domain of the viroid (Keese and Symons, 1985; Visvader and Symons, 1985). This interpretation fits with the fact that up-to-date no HSVd was reported in chickpea and therefore the entity here reported could be a novel HSVd variant.

The data obtained on *CymRSV* indicates that for "known–known" and some "Unknown–known" viruses the ref-seq approach may find practical (cost-effective) applications in particular for surveys of viruses for diagnosis in agro-ecosystems, plant

population (e.g., old varieties) and single plant tissues/organs or for wider environmental studies of ecogenomics (Roossinck, 2011). Indeed, when applying the same pipeline to the sRNA library of a plant population constituting an old Chickpea variety we were able to reveal the presence of *ToMMV* (NC_022230), a putative novel *Tobamovirus* naturally infecting tomatoes in Mexico (Li et al., 2013). Note to worthy, we show through a metagenomics approach that the *ToMMV* is already present in Europe and that in can be hosted by *C. arietinum*. The ToMMV has been just proposed as a novel species of the *Tobamovirus* genus based on sequence similarity with other species of the genus. Phylogenetic analysis shows that *ToMMV* was clustered together with a group of Tobamoviruses mainly infecting solanaceous plants and therefore the presence in chickpea may give good reasons for further characterization of the viral genome, i.e., by generating a higher sRNA sequencing depth with the aim to increase the overall coverage. In cases where plant populations are studied classical molecular approaches are not always applicable. Indeed, an RT-PCR strategy was designed on assembled contigs and attempted on total RNA (see Materials and Methods) but no amplified products were observed. This may be due to the very low titer of virus in the tissues and/or to the infection of a discrete number of plants within the population composing the variety. The presence of a *Tobamovirus* into old varieties of Chickpea in Puglia is not surprising since this genus of plant viruses is known to be hosted by a wide range of plant species, including legumes. Moreover, all viral species are known to be transmitted mechanically and also through contaminations of the seed teguments (Broadbent, 1965).

In summary, a reference-guided approach appeared the most efficient in reconstructing viral metagenomes. Our results indicate that, using an appropriate short-read alignment mapping tool, even low abundant viruses can be well reconstructed. (e.g., at average sequencing depth lower than 7%, still 87% of the virus genome could be assembled). The *de novo* assembly based approaches reached a non-appreciable genome coverage and show a relatively high degree of fragmentation. Nonetheless the contigs generated were sufficiently long for assigning a proper taxonomic classification. Remarkably, the removal of duplicate sequences or the use of Metavelvet assembly software, which is specifically designed for metagenomics applications, did not contribute to more complete assemblies (i.e., longer contigs). At different *k*-mer settings all genome-based assembly strategies used yield similar genome coverage. In contrast, the use of a transcriptome-based method such as Oases resulted in longer contigs and may therefore be the method of choice for v-siRNA-based assemblies. Even if the total genome coverage was lower than in the case of genome-based assembly strategies, the increased average length may provide better anchor points for primer design of PCR products.

Thus, we deduce that transcriptome-based algorithms (i.e., those having adapted the original De Bruijn graph to assemble differentially expressed non-repetitive genes) could better manage the (large) differences between frequencies of sRNA reads covering the various viruses. Strategies based on the original De Bruijn graph algorithm instead attempt to remove *k*-mers with extreme abundance (also simulated by our REMRED approach). Moreover metagenomic-based assembly strategies need to overcome genome-repetitiveness in addition to differences in genome coverage: it may be that the complexity of these issues is higher than the v-siRNA assembly problem where genome repetitiveness is less of an issue.

We conclude that no method or particular *k*-mer setting was able to generate a full coverage, but also that different parameter settings led to assembly of unique (non-overlapping) v-siRNAs. Thus the use of a consensus-based strategy, where a master consensus genome is constructed from multiple assemblies (i.e., at different settings) could potentially be a more robust approach to reconstruct more complete viral genomes.

ACKNOWLEDGMENTS

The work was mainly supported by the project SaVeGrainIN-Puglia – Progetti integrati per la biodiversità "PSR" Regione Puglia FEASR 2007-2013. Reg. (CE) 1698/2005.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00790/abstract

REFERENCES

Aliyari, R., and Ding, S. W. (2009). RNA-based viral immunity initiated by the Dicer family of host immune receptors. *Immunol. Rev.* 227, 176–188. doi: 10.1111/j.1600-065X.2008.00722.x

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. doi: 10.1016/S0022-2836(05)80360-2

Barba, M., Czosnek, H., and Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* 6, 106–136. doi: 10.3390/v6010106

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13:R122. doi: 10.1186/gb-2012-13-12-r122

Broadbent, L. (1965). The epidemiology of TMV. *Ann. Appl. Biol.* 56, 177–205. doi: 10.1111/j.1744-7348.1965.tb01227.x

Candresse, T., Filloux, D., Muhire, B., Julian, C., Galzi, S., Fort, G., et al. (2014). Appearances can be deceptive: revealing a hidden viral infection with deep sequencing in a plant quarantine context. *PLoS ONE* 9:e102945. doi: 10.1371/journal.pone.0102945

De Souza, J., Fuentes, S., Savenkov, E. I., Cuellar, W., and Kreuze, J. F. (2013). The complete nucleotide sequence of sweet potato C6 virus: a carlavirus lacking a cysteine-rich protein. *Arch. Virol.* 158, 1393–1396. doi: 10.1007/s00705-013-1614-x

Ding, S. W., and Voinnet, O. (2007). Antiviral immunity directed by small RNAs. *Cell* 130, 413–426. doi: 10.1016/j.cell.2007.07.039

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. *Microbiol. Mol. Biol. Rev.* 76, 159–216. doi: 10.1128/MMBR.05023-11

Donaire, L., Barajas, D., Martinez-Garcia, B., Martinez-Priego, L., Pagan, I., and Llave, C. (2008). Structural and genetic requirements for the biogenesis of tobacco rattle virus-derived small interfering RNAs. *J. Virol.* 82, 5167–5177. doi: 10.1128/JVI.00272-08

Garcia, D., Garcia, S., Pontier, D., Marchais, A., Renou, J. P., Lagrange, T., et al. (2012). Ago hook and RNA helicase motifs underpin dual roles for SDE3 in antiviral defense and silencing of nonconserved intergenic regions. *Mol. Cell* 48, 109–120. doi: 10.1016/j.molcel.2012.07.028

Garcia-Ruiz, H., Takeda, A., Chapman, E. J., Sullivan, C. M., Fahlgren, N., Brempelis, K. J., et al. (2010). *Arabidopsis* RNA-dependent RNA polymerases and dicer-like proteins in antiviral defense and small interfering RNA biogenesis during Turnip Mosaic Virus infection. *Plant Cell* 22, 481–496. doi: 10.1105/tpc.109.073056

Giampetruzzi, A., Roumi, V., Roberto, R., Malossini, U., Yoshikawa, N., La Notte, P., et al. (2012). A new grapevine virus discovered by deep sequencing of virus-

and viroid-derived small RNAs in Cv Pinot gris. *Virus Res.* 163, 262–268. doi: 10.1016/j.virusres.2011.10.010

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Johansen, L. K., and Carrington, J. C. (2001). Silencing on the spot. Induction and suppression of RNA silencing in the *Agrobacterium*-mediated transient expression system. *Plant Physiol.* 126, 930–938. doi: 10.1104/pp.126.3.930

Keese, P., and Symons, R. H. (1985). Domains in viroids: evidence of intermolecular RNA rearrangements and their contribution to viroid evolution. *Proc. Natl. Acad. Sci. U.S.A.* 82, 4582–4586. doi: 10.1073/pnas.82.14.4582

Kreuze, J., Koenig, R., De Souza, J., Vetten, H. J., Muller, G., Flores, B., et al. (2013). The complete genome sequences of a Peruvian and a Colombian isolate of Andean potato latent virus and partial sequences of further isolates suggest the existence of two distinct potato-infecting tymovirus species. *Virus Res.* 173, 431–435. doi: 10.1016/j.virusres.2013.01.014

Kreuze, J. F., Perez, A., Untiveros, M., Quispe, D., Fuentes, S., Barker, I., et al. (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. doi: 10.1016/j.virol.2009.03.024

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

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Li, R., Gao, S., Fei, Z., and Ling, K. S. (2013). Complete genome sequence of a new *Tobamovirus* naturally infecting tomatoes in Mexico. *Genome Announc.* 1:e00794-13.

Loconsole, G., Saldarelli, P., Doddapaneni, H., Savino, V., Martelli, G. P., and Saponari, M. (2012). Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family Geminiviridae. *Virology* 432, 162–172. doi: 10.1016/j.virol.2012.06.005

MacLean, D., Jones, J. D., and Studholme, D. J. (2009). Application of 'next-generation' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.* 7, 287–296.

Marais, A., Faure, C., Mustafayev, E., Barone, M., Alioto, D., and Candresse, T. (2015). Characterization by deep sequencing of *Prunus* virus T, a novel Tepovirus infecting *Prunus* species. *Phytopathology* 105, 135–140. doi: 10.1094/PHYTO-04-14-0125-R

Massart, S., Olmos, A., Jijakli, H., and Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* 188, 90–96. doi: 10.1016/j.virusres.2014.03.029

Mokili, J. L., Rohwer, F., and Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* 2, 63–77. doi: 10.1016/j.coviro.2011.12.004

Moxon, S., Jing, R., Szittya, G., Schwach, F., Rusholme Pilcher, R. L., Moulton, V., et al. (2008). Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Res.* 18, 1602–1609. doi: 10.1101/gr.080127.108

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40:e155. doi: 10.1093/nar/gks678

Navarro, B., Pantaleo, V., Gisel, A., Moxon, S., Dalmay, T., Bisztray, G., et al. (2009). Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. *PLoS ONE* 4:e7686. doi: 10.1371/journal.pone.0007686

Pantaleo, V., and Burgyan, J. (2008). Cymbidium ringspot virus harnesses RNA silencing to control the accumulation of virus parasite satellite RNA. *J. Virol.* 82, 11851–11858. doi: 10.1128/JVI.01343-08

Pantaleo, V., Saldarelli, P., Miozzi, L., Giampetruzzi, A., Gisel, A., Moxon, S., et al. (2010). Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine. *Virology* 408, 49–56. doi: 10.1016/j.virol.2010.09.001

Pantaleo, V., Szittya, G., and Burgyan, J. (2007). Molecular bases of viral RNA targeting by viral small interfering RNA-programmed RISC. *J. Virol.* 81, 3797–3806. doi: 10.1128/JVI.02383-06

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes Dev.* 20, 3407–3425. doi: 10.1101/gad.1476406

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Roossinck, M. J. (2011). Environmental viruses from biodiversity to ecology. *Curr. Opin. Virol.* 1, 50–51. doi: 10.1016/j.coviro.2011.05.012

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annu. Rev. Genet.* 46, 359–369. doi: 10.1146/annurev-genet-110711-155600

Ruiz-Ferrer, V., and Voinnet, O. (2009). Roles of plant small RNAs in biotic stress responses. *Annu. Rev. Plant Biol.* 60, 485–510. doi: 10.1146/annurev.arplant.043008.092111

Russo, M., Burgyan, J., and Martelli, G. P. (1994). Molecular biology of tombusviridae. *Adv. Virus Res.* 44, 381–428. doi: 10.1016/S0065-3527(08)60334-6

Schloss, L., Falk, K. I., Skoog, E., Brytting, M., Linde, A., and Aurelius, E. (2009). Monitoring of *Herpes Simplex Virus* DNA types 1 and 2 viral load in cerebrospinal fluid by real-time PCR in patients with herpes simplex encephalitis. *J. Med. Virol.* 81, 1432–1437. doi: 10.1002/jmv.21563

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012). Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28, 1086–1092. doi: 10.1093/bioinformatics/bts094

Seguin, J., Rajeswaran, R., Malpica-Lopez, N., Martin, R. R., Kasschau, K., Dolja, V. V., et al. (2014). De novo reconstruction of consensus master genomes of plant RNA and DNA viruses from siRNAs. *PLoS ONE* 9:e88513. doi: 10.1371/journal.pone.0088513

Shimura, H., and Pantaleo, V. (2011). Viral induction and suppression of RNA silencing in plants. *Biochim. Biophys. Acta* 1809, 601–612. doi: 10.1016/j.bbagrm.2011.04.005

Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., et al. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* 3:4. doi: 10.1186/1758-907X-3-4

Stobbe, A. H., and Roossinck, M. J. (2014). Plant virus metagenomics: what we know and why we need to know more. *Front. Plant Sci.* 5:150. doi: 10.3389/fpls.2014.00150

Szittya, G., Moxon, S., Pantaleo, V., Toth, G., Rusholme Pilcher, R. L., Moulton, V., et al. (2010). Structural and functional analysis of viral siRNAs. *PLoS Pathog.* 6:e1000838. doi: 10.1371/journal.ppat.1000838

Vaistij, F. E., and Jones, L. (2009). Compromised virus-induced gene silencing in RDR6-deficient plants. *Plant Physiol.* 149, 1399–1407. doi: 10.1104/pp.108.132688

Visvader, J. E., and Symons, R. H. (1985). Eleven new sequence variants of citrus exocortis viroid and the correlation of sequence with pathogenicity. *Nucleic Acids Res.* 13, 2907–2920. doi: 10.1093/nar/13.8.2907

Wang, X. B., Wu, Q., Ito, T., Cillo, F., Li, W. X., Chen, X., et al. (2010). RNAi-mediated viral immunity requires amplification of virus-derived siRNAs in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 107, 484–489. doi: 10.1073/pnas.0904086107

Wassenegger, M., and Krczal, G. (2006). Nomenclature and functions of RNA-directed RNA polymerases. *Trends Plant Sci.* 11, 142–151. doi: 10.1016/j.tplants.2006.01.003

Wu, Q., Luo, Y., Lu, R., Lau, N., Lai, E. C., Li, W. X., et al. (2010). Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs. *Proc. Natl. Acad. Sci. U.S.A.* 107, 1606–1611. doi: 10.1073/pnas.0911353107

Wu, Q., Wang, Y., Cao, M., Pantaleo, V., Burgyan, J., Li, W. X., et al. (2012). Homology-independent discovery of replicating pathogenic circular RNAs by

deep sequencing and a new computational algorithm. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3938–3943. doi: 10.1073/pnas.1117815109

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

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

Received: 06 November 2014; paper pending published: 25 November 2014; accepted: 22 December 2014; published online: 27 January 2015.

*Citation: Pirovano W, Miozzi L, Boetzer M and Pantaleo V (2015) Bioinformatics approaches for viral metagenomics in plants using short RNAs: model case of study and application to a Cicer arietinum population. Front. Microbiol. **5**:790. doi: 10.3389/fmicb.2014.00790*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 04 March 2016
doi: 10.3389/fmicb.2016.00269

GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in *Alpavirinae* Viral Discovery from Metagenomic Data

João M. P. Alves[1†], André L. de Oliveira[1†], Tatiana O. M. Sandberg[1], Jaime L. Moreno-Gallego[2], Marcelo A. F. de Toledo[1], Elisabeth M. M. de Moura[3], Liliane S. Oliveira[1,4], Alan M. Durham[4], Dolores U. Mehnert[3], Paolo M. de A. Zanotto[3], Alejandro Reyes[5,6] and Arthur Gruber[1]**

[1] Department of Parasitology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil, [2] Graduate program in Computational Biology, Universidad de los Andes, Bogotá, Colombia, [3] Department of Microbiology, Institute of Biomedical Sciences, University of São Paulo, São Paulo, Brazil, [4] Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil, [5] Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia, [6] Center for Genome Sciences and Systems Biology, Department of Pathology and Immunology, Washington University in Saint Louis, MO, USA

This work reports the development of GenSeed-HMM, a program that implements seed-driven progressive assembly, an approach to reconstruct specific sequences from unassembled data, starting from short nucleotide or protein seed sequences or profile Hidden Markov Models (HMM). The program can use any one of a number of sequence assemblers. Assembly is performed in multiple steps and relatively few reads are used in each cycle, consequently the program demands low computational resources. As a proof-of-concept and to demonstrate the power of HMM-driven progressive assemblies, GenSeed-HMM was applied to metagenomic datasets in the search for diverse ssDNA bacteriophages from the recently described *Alpavirinae* subfamily. Profile HMMs were built using *Alpavirinae*-specific regions from multiple sequence alignments (MSA) using either the viral protein 1 (VP1; major capsid protein) or VP4 (genome replication initiation protein). These profile HMMs were used by GenSeed-HMM (running Newbler assembler) as seeds to reconstruct viral genomes from sequencing datasets of human fecal samples. All contigs obtained were annotated and taxonomically classified using similarity searches and phylogenetic analyses. The most specific profile HMM seed enabled the reconstruction of 45 partial or complete *Alpavirinae* genomic sequences. A comparison with conventional (global) assembly of the same original dataset, using Newbler in a standalone execution, revealed that GenSeed-HMM outperformed global genomic assembly in several metrics employed. This approach is capable of detecting organisms that have not been used in the construction of the profile HMM, which opens up the possibility of diagnosing novel viruses, without previous specific information, constituting a *de novo* diagnosis. Additional applications include, but are not limited to, the specific assembly of extrachromosomal elements such as plastid and mitochondrial genomes

OPEN ACCESS

Edited by:
Akio Adachi,
Tokushima University Graduate School, Japan

Reviewed by:
Thierry Candresse,
Institut National de la Recherche Agronomique, France
Makoto Kuroda,
National Institute of Infectious Diseases, Japan

***Correspondence:**
Alejandro Reyes
a.reyes@uniandes.edu.co;
Arthur Gruber
argruber@usp.br

[†] *These authors have contributed equally to the work.*

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *29 October 2015*
Accepted: *19 February 2016*
Published: *04 March 2016*

Citation:
Alves JMP, de Oliveira AL, Sandberg TOM, Moreno-Gallego JL, de Toledo MAF, de Moura EMM, Oliveira LS, Durham AM, Mehnert DU, Zanotto PMA, Reyes A and Gruber A (2016) GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data. Front. Microbiol. 7:269. doi: 10.3389/fmicb.2016.00269

from metagenomic data. Profile HMM seeds can also be used to reconstruct specific protein coding genes for gene diversity studies, and to determine all possible gene variants present in a metagenomic sample. Such surveys could be useful to detect the emergence of drug-resistance variants in sensitive environments such as hospitals and animal production facilities, where antibiotics are regularly used. Finally, GenSeed-HMM can be used as an adjunct for gap closure on assembly finishing projects, by using multiple contig ends as anchored seeds.

Keywords: *Alpavirinae*, sequence assembly, metagenomic analysis, viral discovery, *de novo* diagnosis

INTRODUCTION

From the golden age of phage research establishing the basis for the development of molecular biology, virus research suffered a decline due to several technical difficulties, in particular the necessity of knowing the specific viral and host life cycles and conditions for *in vitro* growth (Rosenberg, 2015). With the advent of next generation sequencing (NGS) and metagenomics, viral discovery and research entered a new successful age. A pioneering metagenome study, a virome of uncultured marine viral communities (Breitbart et al., 2002), revealed a predominance of bacteriophages, and demonstrated the potential of metagenomics in the field of viral research. Since then, viral ecology has risen as a new field, and it is now possible to assess the viral composition of a microbial community and understand the fundamental role that these highly abundant biological entities play in any environment, with particular efforts shown in marine environments (Rohwer and Thurber, 2009). However, since the very start of the metagenomic bloom, it has been clear that our knowledge of viral diversity is scarce and relies on viruses where the host is known and can be cultivated, severely restricting the known viral diversity to possibly less than 1% of what is actually out there (*cf.* Fancello et al., 2012). Furthermore, the rate of shotgun data generation has outpaced the sequencing of reference viral genomes, and this ever-increasing gap limits our capacity to analyze newly generated datasets. Thus, the development of new computational tools is of utmost importance to increase our understanding of viral diversity (Fancello et al., 2012). Some of the most important pandemic diseases arose by the transmission of viruses originally present in animals that were able to adapt to the human host (Wang, 2011; Rosenberg, 2015). Thus, a systematic surveillance for emerging viruses is crucial to enable the detection of novel and potentially devastating ones before they become pandemic (Lipkin and Firth, 2013; Smits and Osterhaus, 2013).

The human and animal microbiome field has benefited immensely from the advances in NGS and metagenomics (Tang and Chiu, 2010; Bexfield and Kellam, 2011). The number of studies characterizing the gut microbiome has increased exponentially in recent years, and such studies have linked changes in these complex communities to diseases ranging from obesity and malnutrition to even Alzheimer's and autism (Mayer et al., 2014). An important component of this microbial community is the viral one, in particular phages that are an integral part of the community (Reyes et al., 2010, 2012, 2015; Dutilh et al., 2014; Norman et al., 2015). Since the early studies of the viral component of the gut microbiota, an important limitation has been the lack of reference viral genomes infecting the Firmicutes and Bacteroidetes, which constitute the most abundant bacterial phyla inhabiting the gut (Arumugam et al., 2011) Bacteriophages are gaining growing relevance in gut microbiome studies where changes in viral and phage population have been linked to alterations in the microbial community and/or human health (Norman et al., 2015; Reyes et al., 2015).

Alpavirinae, a recently characterized subfamily of the *Microviridae* family, is composed of ssDNA phages that exist either as temperate phages of Bacteroidetes genomes (Kim et al., 2011; Krupovic and Forterre, 2011) or infectious particles (Roux et al., 2012; Zhong et al., 2015). Roux et al. (2012) analyzed metagenomic data from different geographic locations and biological sources, and described a large set of complete, previously undescribed *Microviridae* genomes, including 33 *Alpavirinae* genomes. More recently, Quaiser et al. (2015) described 17 additional complete *Microviridae* genomes from a *Sphagnum*-dominated peatland. A recent study (Zhong et al., 2015) reported the occurrence of *Microviridae* in peri-alpine lakes, mainly represented by gokushoviruses, but also including *Alpavirinae*, a finding that confirms that this latter group is also present in fresh waters, possibly in both lysogenic and lytic forms. Cantalupo et al. (2011) found diverse viral populations in raw sewage, with 80% of the metagenomic reads being related to bacteriophages and, from this subset, 37% were derived from *Microviridae*. Considering that relatively few genomes of the *Alpavirinae* subfamily have been described so far and their initial description as Bacteroidetes associated viruses, this taxonomic group constitutes an interesting case study for a new viral discovery strategy.

One of the most challenging tasks for metagenomic data analysis is the assembly phase (Wajid and Serpedin, 2012; El-Metwally et al., 2013). Several algorithms have been developed and can roughly be classified according to the graph construction method: greedy, OLC (overlap-layout-consensus), and de Bruijn graphs. Assemblers using the OLC method are most appropriate for datasets of relatively long reads, such as Sanger and 454 platforms, but the quadratic complexity of the overlap computation phase severely limits the size of the datasets that can be used. Assemblers using *k*-mers and de Bruijn graphs require much less computational power, but memory requirement is still high. Therefore, whatever the algorithm, sequence assemblers are highly demanding in terms of memory

usage and/or processing power, especially for datasets in the magnitude of millions of reads. Additionally, most *de novo* assemblers have been developed for single-organism genome sequencing (Fancello et al., 2012). In fact, *de novo* assembly of metagenomic data is particularly challenging for several reasons, among others: (1) the heterogeneous nature of the sample, with many different organisms; (2) uneven distribution of organism quantities, leading to biased sampling and coverage; (3) unlike single-organism genome sequencing, the number of final assembled sequences cannot be predicted; (4) sequences derived from closely related organisms may generate chimeric assemblies; (5) polymorphisms, in a way similar to sequencing errors, can disrupt assemblies by tangling the assembly graph (i.e., by creating specific topological structures such as tips and bubbles). With those challenges in mind, a few recent attempts have been made to either modify traditional assemblers or develop assemblers specifically designed for metagenomic data (Fancello et al., 2012). However, such approaches still suffer from the same computational resource drawbacks mentioned above for traditional genome assemblers.

Many sequencing projects do not have as a goal the reconstruction of all possible sequences present in a sample, but rather aim at studying a well-defined gene, gene family, or a transcript. In this case, a target-specific assembly could represent a more sensible approach. To fulfill such a need, our group was the first one to develop a seed-driven progressive assembly algorithm, implemented in the GenSeed program (Sobreira and Gruber, 2008), as a rational method to reconstruct specific targets from unassembled sequence datasets. GenSeed uses a short DNA or protein sequence as a query in similarity searches to select reads, which in turn are retrieved from the dataset and assembled together with the seed sequence, leading to an increment of its original length. Short sequences are then extracted from the assembled sequence ends and used as new seeds in an iterative process that generates progressively longer sequences at each assembly cycle. Because assembly is performed in multiple steps and relatively few reads are used in each cycle, the program demands low computational resources. Some recent approaches based on the same concept of seed-driven iterative assembly have been proposed for the assembly of viral sequences from metagenomic data (Smits et al., 2015), but they are all restricted to the use of DNA sequence seeds. In this work, we report the development of GenSeed-HMM, a completely revised and highly incremented version of GenSeed. The proposed approach relies on two principles: (1) progressive assembly as an alternative for sequence reconstruction; and (2) the use of profile HMMs as starting seeds for target-driven reconstruction. As a proof of principle, we use GenSeed-HMM and profile HMMs built from *Alpavirinae* proteins to reconstruct novel viral sequences from human fecal samples. GenSeed-HMM allowed the reconstruction of many *Alpavirinae* genomes distinguishable from those described by Roux et al. (2012), outperforming conventional (global) genomic assembly in several metrics. GenSeed-HMM provides a fast and simple way to run progressive sequence assembly pipelines that are directly targeted at sequences of interest, potentially detecting members of a taxonomic group related but not equal to those used on the construction of the profile HMM. This feature opens up the possibility of diagnosing novel viruses, without previous specific information.

MATERIALS AND METHODS

Data Sources

Two distinct metagenomic datasets were used in this study, derived from fecal microbiota and raw sewage samples. The metagenomic sequence data from fecal microbiota was obtained from monozygotic twins and their mothers, and sequenced on the 454 platform, as previously described (Reyes et al., 2010). Sequence datasets (accession codes SRX028823 to SRX028827) were downloaded from the Sequence Read Archive (SRA) at http://www.ncbi.nlm.nih.gov/sra. SRA format files were converted into FASTQ using the *fastq-dump* program (SRA toolkit) and all adaptors were trimmed with *cutadapt* (https://cutadapt.readthedocs.org) using parameters `–q 30 –minimum–length 50 –overlap=5 –u 14`. Raw sewage (total volume of 15 L) collected at the municipality of Taboão da Serra (São Paulo, Brazil) was pressure-filtered through an AP-20 filter membrane (Merck Millipore) and electropositive filter membranes Zeta Plus 60 (AMF, Cuno Div.). Viruses were then eluted in a protein mix, concentrated by ultracentrifugation and treated with Vertrel XF (decafluoropentane, DuPont) to remove lipids and proteins (Mehnert and Stewien, 1993; Queiroz et al., 2001). Viral DNA was extracted using DNeasy Blood and Tissue kit (Qiagen®) and amplified with an illustra™Single Cell GenomiPhi™DNA Amplification Kit (GE Healthcare Life Sciences). The DNA was used to construct a library with the Nextera XT DNA Library Preparation Kit (Illumina, Inc.) and sequenced using the Illumina HiSeq 2500 System, generating 101-bp paired-end reads. To remove the Nextera transposase sequence, FASTQ files were trimmed with *cutadapt* using parameters `–q 30 –a CTGTCTCTTATACACATCT –minimum–length 50 –overlap=5 –u 2`.

GenSeed-HMM Development and Progressive Assembly

GenSeed-HMM was developed in the Perl language and is publicly available for download under the terms of the GNU General Public License version 3 at http://genseedhmm.sourceforge.net. Installation instructions and documentation are also provided. All tests reported in this work were performed on a Dell PowerEdge T710 server with two Intel Xeon X5660 2.8 Ghz processors and 64 GB of RAM. GenSeed-HMM can be used in any POSIX-compliant operating system such as UNIX and Linux distributions with an installed Perl interpreter (http://www.perl.org). The list of programs required by GenSeed-HMM varies according to the type of seed employed and the assembler that will be used, as well as whether mapping of recruited reads to resulting contigs is desired. For profile HMM seeds, the following packages/programs are required: *transeq* from the EMBOSS package (Rice et al., 2000), BLAST+ (Camacho et al., 2009), and HMMER v3.0 (Eddy, 2011). For the optional mapping of

recruited reads against resulting contigs, Bowtie2 (Langmead and Salzberg, 2012). GenSeed-HMM requires at least one installed DNA assembler and is compatible with the following programs: SOAPdenovo (Luo et al., 2012), ABySS (Simpson et al., 2009), Velvet (Zerbino and Birney, 2008), Newbler (GS *De Novo* Assembler, Roche 454 Life Sciences, available under request at http://my454.com/contact-us/software-request.asp), and CAP3 (Huang and Madan, 1999). If Newbler is to be used, programs *sfffile* and *sffinfo* (both distributed by Roche 454 Life Sciences) and *splitter* (from EMBOSS) are also required. Progressive assemblies were performed using GenSeed-HMM. Several parameter sets were tested to optimize assembly results. Parameters used in the final experiments reported here are: `–assembler newbler –ext_seed_size 30 –max_contig_length 10000 –threads 20 –clean no –mapping yes –blastn_parameters "–evalue 0.0001 –num_threads 20 –dust no –perc_identity 85" –no_qual`. Specific profile HMM seeds (Supplementary File 1) were used throughout this work and specified on GenSeed-HMM with parameter `–seed`.

Profile HMM Construction

For profile reconstruction, all available sequences corresponding to previously reported viral proteins (VP) of *Alpavirinae* (Roux et al., 2012), named VP1, VP2, VP3, and VP4, were retrieved. Multiple sequence alignments (MSA) of each group of proteins were created using MUSCLE (Edgar, 2004) with default parameters, and the alignments were manually inspected with Jalview (Waterhouse et al., 2009) to identify conserved regions. The MSA was appended with the respective (VP1, VP2, VP3, or VP4) proteins from *Gokushovirinae* and *Pichovirinae* in order to determine whether identified conserved regions were subfamily specific. Specific regions on VP1 (Supplementary Figure 1) and VP4 (not shown) were selected and profile HMMs were built using *hmmbuild* from the HMMER package (Eddy, 2011). We adopted a nomenclature composed of the viral protein name (e.g., VP1) plus the region (e.g., R4) of the multiple sequence alignment chosen to build the respective profile HMM used as seed.

Assembly Evaluation and Cross-Similarity Analysis of Contigs Reconstructed with Different Profile HMM Seeds

Contigs assembled with GenSeed-HMM were analyzed with in-house scripts to list and calculate contig lengths and generate contig size ranks. Contigs reconstructed by progressive assembly using GenSeed-HMM with different profile HMM seeds were sorted in descending order by length and submitted to an all-vs-all *blastn* similarity search. Clusters included contigs presenting at least 90% similarity at the nucleotide level, covering at least 90% of the length of the shortest contig. Contig clusters were used to evaluate consistency between assemblies based on different profile HMM seeds to identify the potential minimum contig set.

Taxonomic Assignment of Contigs

For taxonomic assignment of assembled contigs, *blastx* similarity search was used to compare assembled contigs against all reference *Microviridae* proteins (Roux et al., 2012) with a cutoff *E*-value of 1e-20. The top 10 hits were manually checked for consistency and taxonomic assignment was given to the subfamily to which all significant hits were observed. Taxonomic assignment was set to all subfamilies matched in cases where hits with similar scores were obtained to more than one subfamily. Taxonomic assignment to each cluster was done by comparing individual contig assignments within each cluster; for all clusters, we observed 100% agreement in taxonomic classification among the contigs constituting the corresponding cluster.

Contig Distribution from Different Human Samples

Sequence reads derived from each human donor fecal sample (Reyes et al., 2010) were mapped using Bowtie2 (Langmead and Salzberg, 2012) to the assembled contigs assembled by GenSeed-HMM using the VP1R4 seed. Mapping counts were normalized by contig length and sample sequencing effort (RPKM—Reads Per Kilobase per Million mapped reads), and log transformed. The resulting matrix was used to generate a heatmap diagram.

Sequence Analysis and Annotation

All assembled contigs were submitted to an automatic annotation pipeline using the development version EGene 2, derived from the EGene platform (Durham et al., 2005). The pipeline starts with a gene prediction step using Glimmer 3.02 (Delcher et al., 2007) using a training set composed of *Alpavirinae* proteins (Roux et al., 2012). All translated products were then submitted to *blastp* searches against the non-redundant (nr) database and a database composed of proteins derived from *Microviridae*. Hits were considered positive when presenting *E*-values below 1e-6. Protein domains and families were subsequently identified via InterPro (Mitchell et al., 2015) searches. In the specific case of contig annotation from the VP1R4 assembly, annotation has been manually curated to find missing and/or truncated ORFs. Automatic annotations and all stored evidence for contigs are publicly available at http://www.coccidia.icb.usp.br/alpavirinae.

Phylogenetic Analysis

For each contig assembled using the VP1R4 profile HMM seed, the complete or partial VP1 sequence was identified, translated and used for phylogenetic analyses. Two sets of analyses were done: one using only complete VP1 proteins, while the other used only a conserved region present in all assembled contigs consisting of approximately 75 amino acids having the VP1R4 region at the C-terminus. Each set of proteins was complemented with reference VP1 proteins from published datasets (Roux et al., 2012) and GenBank-deposited datasets (see Supplementary Table 1) belonging to other *Microviridae* subfamilies: *Gokushovirinae*, *Pichovirinae,* and genus *Microvirus*. Protein alignments were performed using MUSCLE (Edgar, 2004) and manually edited using Jalview (Waterhouse et al., 2009). Phylogenetic analyses were performed using maximum-likelihood (ML) in RAxML 8.2.0 (Stamatakis, 2006) The best-fitting amino acid substitution model for each set was obtained with ProtTest 3.4 using the AIC

statistic for model selection (Darriba et al., 2011). Finally, support for nodes in ML trees was assessed by bootstrap analysis with 100 pseudoreplicates and support values were added to the master ML tree.

Comparison of Progressive vs. Global Assembly

To compare progressive assemblies with the global assembly counterparts, we ran Newbler as a standalone application, with default parameters, using the complete read datasets for single-end 454 (human fecal samples) and paired-end Illumina (sewage samples) data. For the latter, assembly was performed taking into account paired-end information in order to generate the best possible global assembly. All contig sequences obtained were translated into the six possible reading frames using *transeq* and then used as a dataset for *hmmsearch* (HMMER3 package) using the VP1R4 profile HMM as query. Contigs coding for HMM-positive protein sequences were identified and their nucleotide sequences used for size ranking and comparison to contigs assembled by GenSeed-HMM.

Coverage Analysis

Read alignment (SAM) files produced by GenSeed-HMM were loaded onto Tablet (Milne et al., 2013; https://ics.hutton.ac.uk/tablet/) and used to generate base-by-base coverage files for each assembled contig. Coverage information of global assembly was obtained from alignment information files produced by Newbler, and average per-base-coverage for each contig was calculated. VP1R4-containing contigs, derived from the global and progressive assembly, were pooled together and submitted to a *blastn* all-vs-all similarity search. Contigs that were at least 97% identical at the nucleotide level over at least 90% of the length of the shortest contig were clustered.

RESULTS

GenSeed-HMM Implementation and Execution

GenSeed-HMM is a completely revised and extended version of the previously described GenSeed program (Sobreira and Gruber, 2008). With the advent of next-generation sequencing (NGS) platforms, the ability to use up-to-date sequencing data and DNA assemblers became an essential feature for any sequence reconstruction program. Hence, several improvements over GenSeed's original implementation have been implemented: (1) in addition to CAP3, GenSeed-HMM can now use Newbler, Velvet, SOAPdenovo, or ABySS as third-party assemblers; (2) input formats now include FASTA, FASTA.QUAL, FASTQ, and SFF, including the possibility of using quality values for CAP3 and Newbler; (3) instead of BLAST, GenSeed-HMM now uses BLAST+, a new version of the BLAST suite that uses the NCBI C++ Toolkit and presents several performance and feature improvements; and (4), in addition to DNA and protein sequences, profile HMMs can now be employed as seeds by using HMMER3, a package that performs similarity searches using profile HMMs as queries, with a performance comparable to BLAST. GenSeed-HMM automatically detects seed type (DNA, protein or profile HMM; **Figure 1A**). The program accepts as input a sequencing dataset generated by any of a variety of platforms and, in our experience, GenSeed-HMM can effectively reconstruct sequences using datasets originating from Sanger, 454, or Illumina technologies, with reads as short as 35-bp (data not shown). The dataset format is automatically identified and, if necessary, converted to FASTA. The database for BLAST+ is then generated by *makeblastdb* (from the BLAST+ package). If a profile HMM is used as a seed, the sequencing dataset is submitted to a six-frame translation using *transeq* (from the EMBOSS package). GenSeed-HMM performs these steps only once and reuses previously generated files in subsequent runs (**Figure 1B**). The progressive assembly cycle (**Figure 1C**) starts either with a similarity search (*blastn* for DNA seeds, *tblastn* for protein seeds) or with a profile search (*hmmsearch* for profile HMM seeds) against the translated sequencing dataset. Whatever the type of similarity search, a list of hits is obtained and used to retrieve all positive reads (and, if applicable, their quality scores) using internal sequence indexer and retriever functions. The reads are then assembled and the contig ends are used as nucleotide seeds for the subsequent assembly round. These sequences, called extension seeds, can have a variable user-defined length compared to the original seed. All assembly steps use the recruited reads combined with the contig sequence from the previous round, to guarantee that previously obtained sequences will not be disrupted by the incorporation of new reads. The use of multiple seeds is implemented in GenSeed-HMM and if two or more growing contigs overlap at a given assembly cycle, the assembler merges them into a newly generated contig. At any cycle there are checkpoints that determine if new reads have been recruited since the last round and if the resulting contigs increased in length compared to the previous round. The progressive assembly process is interrupted if any one of four conditions is satisfied: (1) the contig has reached the optional user-defined maximum length; (2) the optional user-defined number of assembly iterations has been reached; (3) no new read has been recruited by the current extension seeds compared to the preceding round; or (4) no sequence length increment has been observed since the previous round. In this latter case, GenSeed-HMM executes an iterative trimming routine, which may help overcome extension halts caused by sequencing errors. Briefly, the program iteratively trims the ends of the contig, removing an amount of bases corresponding to 25% of the extension seed length at a time (for a maximum of three steps), and tries to repeat the assembly after each trimming phase. If any step succeeds at recruiting new reads and increasing the contig length, the progressive assembly process is resumed. Conversely, the assembly process is finished and GenSeed-HMM proceeds to the final processing and file storing routines (**Figure 1D**). At the final checking procedure, all contigs assembled at the last round are checked for the presence of the original seed, with only seed-positive contigs being stored. Several processing files, including those generated in the intermediate assembly steps can be stored if specified by the user. Since the assembly is progressively generated, no true assembly files (e.g., those listing meaningful contig qualities,

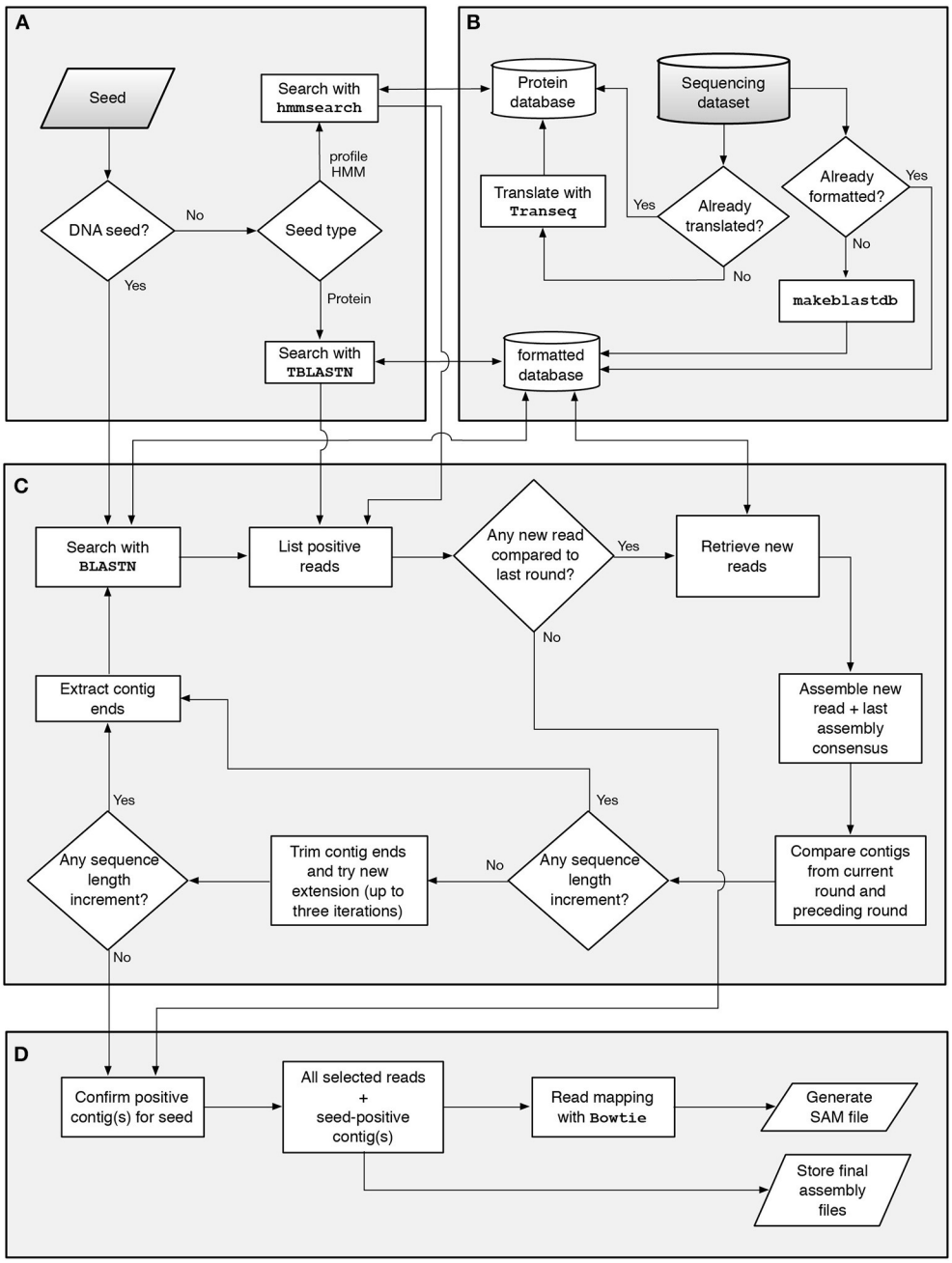


FIGURE 1 | Workflow of the seed-driven progressive assembly process. GenSeed-HMM automatically identifies the type of starting seed **(A)**. The sequencing read database is indexed and, if needed, translated **(B)**. DNA, protein or profile HMM seeds are then used to select reads from the database using *blastn*, *tblastn,* or *hmmsearch*, respectively. The list of positive reads is introduced into the progressive assembly cycle **(C)**. The reads retrieved from the database are assembled and the contig ends are extracted and used as new seeds in an iterative process. The progressive assembly contains several checkpoints and is completed when a set of finishing criteria are fulfilled. In the final procedure **(D)**, all contigs are checked in regard to the presence of the starting seed and final files are stored.

graph information, etc.) are produced. Thus, if required by the user, GenSeed-HMM invokes *bowtie2* to map all recruited reads onto the final contigs. A SAM file is then generated and stored, and can be inspected using a graphical viewer for sequence assemblies and alignments such as *tablet* (Milne et al., 2013).

Profile HMM Design and Use in Progressive Assembly

Since evolutionary processes may impose different selection pressures, proteins may evolve at different rates and even specific domains can present different degrees of conservation. We used GenSeed-HMM in order to identify potential, previously

unidentified viruses belonging to the *Alpavirinae* subfamily, recently identified as part of human gut microbial communities. After analyzing the conservation of the *Microviridae* VP1, VP2, VP3, and VP4 proteins, we decided to initially use a dataset of VP1 and VP4 proteins from available *Alpavirinae* assembled genomes (Roux et al., 2012) to identify conserved regions and then, by appending homologous proteins from *Pichovirinae* and *Gokushovirinae*, select regions with specificity to the subfamily *Alpavirinae*. VP1 is the major capsid protein, a highly conserved protein that has been used as a phylogenetic marker of the group, while VP4 is a genome replication initiation protein and is more diverse in sequence than VP1 (Roux et al., 2012). A total of four distinct regions were selected for each of VP1 (Supplementary Figure 1) and VP4 (not shown) proteins. All profile HMMs were independently tested as seeds in progressive assembly assays using GenSeed-HMM and a dataset derived from viral-like particle (VLP) purification from human fecal samples (Reyes et al., 2010). This dataset is composed of approximately 1.2 million reads generated on the 454 platform, presenting a post-trimming average size of 256 bases. Initially, all contig sets were evaluated by a simple quantitative criterion, considering solely the contig size rank. In the case of VP1 (**Figure 2A**), the VP1R1 profile HMM showed the best performance, with the largest number of long contigs, followed by VP1R4, VP1R5, and VP1R6, respectively. For the VP4 protein (**Figure 2B**), VP4R1, and VP4R3 showed the best results, with VP4R4 and VP4R2 clearly resulting in a much lower number of long-sized contigs. To check the robustness of the method to different NGS technologies, the same profile HMMs were also tested as seeds with a metagenomic dataset derived from raw sewage, composed of 53.5 million Illumina paired-end reads with a post-trimming average size of 92 bases. Either with VP1 (Supplementary Figure 2A) or VP4 (Supplementary Figure 2B) profile HMM seeds, the results were very similar to those observed with fecal samples, with VP1R1, VP4R1, and VP4R3 generating longer contigs than the other seeds.

The variability in the sequence reconstruction ability by the different HMM seeds led us to investigate how the different assemblies compared to each other in terms of their contig sequences. Thus, we used the top four performers (VP1R1, VP1R4, VP4R1, and VP4R3) to run *blastn* all-vs-all similarity searches followed by sequence clustering. It is noteworthy that despite the overall contig size rank variation (**Figures 2A,B**), from the total of 85 de-replicated contigs, 25 were identified as being assembled independently by all four assemblies, whereas using either VP1 or VP4 seeds showed the second highest overlap (**Figure 3**). Therefore, highlighting that regardless of the differences observed in contig size rankings, the assemblies were highly consistent, even though they were derived from profile HMMs built from distinct regions and/or proteins. The only other overlap with a significant number of contigs involved nine contigs shared between VP1R1, VP4R3, and VP4R1. However, further taxonomic assignment (see section below) showed that only two of these contigs were assigned to *Alpavirinae*, suggesting lower precision for these seeds.

A further analysis of assembly performance, in particular regarding to VP1R4 seed, showed that some contigs covering this region have not been assembled using the corresponding seed, but rather by one or more other seeds (**Figure 3**). For instance, 14 contigs were assembled exclusively by the VP1R1 seed, with 10 of them being assigned to the *Alpavirinae* subfamily, and four of those covering the VP1R4 region. A detailed analysis of these latter contigs confirmed that the VP1 proteins of this subset were too divergent to be detected by the VP1R4 seed. This phenomenon was observed in all contigs assigned to *Alpavirinae*, but not assembled by the VP1R4 seed (**Figure 3**—represented by black numbers). Another interesting observation was the fact that six contigs assigned to *Alpavirinae*, and containing the VP1R4 seed, have not been assembled when using this particular seed (**Figure 3**—represented by red numbers). In this case, we identified three events where the VP1R4 seed successfully detected the corresponding reads, but due to a very low coverage on this specific region, Newbler was unable to generate an assembled contig in the first assembly cycle. Finally, for the remaining three events, we identified short sequences on the VP1R4 set of contigs that were very similar (but slightly below our 90% threshold) to contigs assembled by the other seeds. By comparing the read coverage of the shorter contigs with their longer counterparts, it became clear that the ends of the shorter contigs presented lower coverage than the corresponding regions

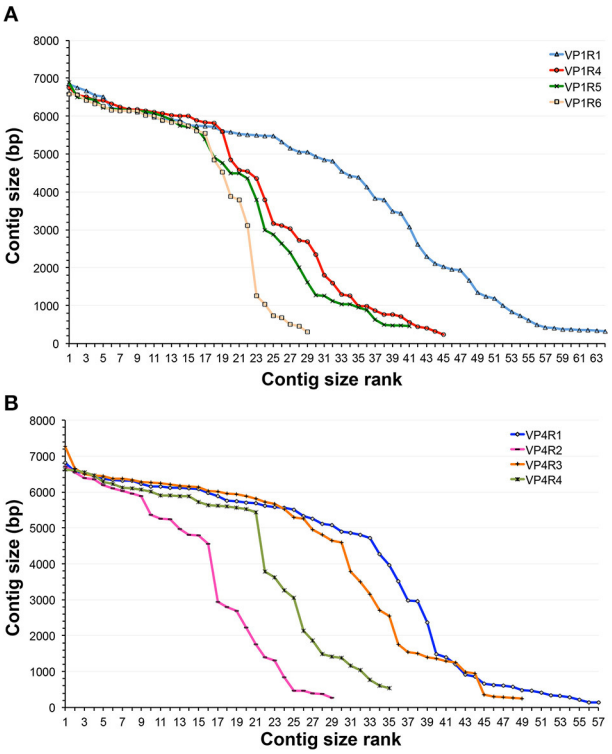


FIGURE 2 | Comparison of progressive assembly using different HMM seeds. Contig profiles obtained by progressive assembly with GenSeed-HMM using a 454 dataset from fecal samples from human patients (Reyes et al., 2010) and profile HMM seeds derived from *Alpavirinae* major capsid protein VP1 **(A)** and replication initiation protein VP4 **(B)**. Contigs are ranked in decreasing order of size. Each marker represents a distinct contig. Profile HMMs used as seeds are depicted.

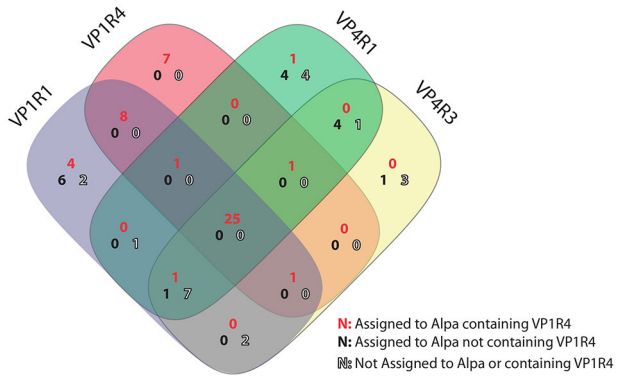


FIGURE 3 | Consistency among HMM seeds. Venn diagram representing shared contigs reconstructed by progressive assembly using GenSeed-HMM with profile HMM seeds VP1R1, VP1R4, VP4R1, and VP4R3. Contigs were included in the same cluster when presenting at least 90% similarity at the nucleotide level covering at least 90% of the length of the shortest contig. Contigs were then taxonomically classified by *blastx* to reference proteins from *Microviridae* and searched for the presence of the VP1R4 seed using *hmmsearch*. A large percent of shared contigs among all four seeds is observed and belonging to *Alpavirinae* genomes covering the VP1R4 seed. Notice that contigs that were not present within the VP1R4 seed were usually not assigned to *Alpavirinae* (low precision) or do not contain the VP1R4 region, suggesting potential shorter non-overlapping contigs.

in the longer ones, indicating premature extension stoppage events (data not shown). This seems to be a consequence of the directionality of the progressive assembly method. The assembler is able to extend the growing sequence in one direction, but, due to base discrepancies biased at a particular end of one or more reads, the resulting alignment graph ends up containing a so-called bubble, precluding the assembler from extending the sequence in the opposite direction. By precisely identifying the few different assembly failures, we expect to develop new routines that could automatically handle these problems, should they happen, during an execution.

Taxonomic Assignment of Assembled Sequences

Given that the aim was to reconstruct *Alpavirinae* genomes from metagenomic datasets, we wanted to address the sensitivity and precision of the methodology. The sensitivity (number of *Alpavirinae* associated contigs from the total number of *Alpavirinae* viruses in a given dataset) and precision (number of *Alpavirinae* associated contigs from the total number of contigs assembled with a given seed) will be dependent on the specific profile HMM seed used, the quality and coverage of the sequencing and the specific parameters used. To address this point we used similarity searches with *blastx* against a reference dataset of *Microviridae* proteins (Roux et al., 2012), together with sequence clustering analysis, and we were able to classify the contigs into three subfamilies of *Microviridae* (**Table 1**). Since all HMMs have been originally built toward *Alpavirinae*-conserved regions, a predominance of sequence assignment to this subfamily was expected. In fact, this was the most prevalent taxon of the reconstructed sequences for all seeds. However, with the exception of VP1R4, which presented 100% precision, the three remainder HMMs also led to assembled *Gokushovirinae* and *Pichovirinae* sequences, with precision values varying from 72.3 to 79.7% (**Table 1**). The unambiguous taxonomic assignment of VP1R4-derived contigs was confirmed by phylogenetic analysis (see below). Clustering analysis showed that among the four assemblies it was possible to generate a total of 85 non-redundant non-overlapping contigs (**Table 1**). However, this result does not necessarily imply that there is a total of 85 different originating viral entities in the sample, since each assembly resulted in a number of partial, shorter contigs centered on the specific profile HMM seed that could be generated from the same virotype but, due to sequencing coverage or other factors affecting assembly, were not extended enough to identify overlaps with contigs produced by other profile HMMs.

Assessing the sensitivity of the different seeds constitutes a challenge since it is impossible to address the real total number of expected *Alpavirinae* genomes. In order to have an approximation to this value we analyzed two different metrics that should constitute an approximate range of the actual sensitivity. As an upper bound, we used the number of total contigs (independently of the seed used) assigned to *Alpavirinae* ($n = 65$; **Table 1**), which is very likely to give an over-estimated sensitivity value due to independent contigs formed by different seeds that originate from a single viral entity, as mentioned above. The lower bound was done specifically for the VP1R4 seed and consists of the number of contigs from all assemblies that covered the region used to build the VP1R4 HMM ($n = 49$; **Table 1**). By the estimation of these contig numbers it was possible to calculate that the sensitivity for the VP1R4-based assembly should be between 66.2 and 87.8% (**Table 1** and **Figure 3**). In a similar way, we calculated the sensitivity and precision of the progressive assembly performed on the sewage data (Supplementary Table 2) in this case we observed for the VP1R4 seed a similar precision (99.67%) and a sensitivity between (35.8–91.6%), the wider range is due to the higher number of total contigs (2480) due to the larger dataset with shorter reads generating a more fragmented assembly.

The advantage of using profile HMMs as seeds for progressive assembly is clear when the same data is investigated by protein similarity searches. With that aim, each of the 33 full-length VP1 sequences from Roux et al. (2012) was compared by *blastp* similarity searches (Supplementary Table 3) to our 45 complete or partial VP1 sequences originated from the VP1R4 assembly. Even using an *E*-value of 1e-6, which is not particularly stringent, we have found that each of the 33 proteins matched only 16–37 of the 45 novel sequences. This shows that a single profile HMM seed derived from a short VP1 region was much more sensitive than any of the 33 complete protein sequences for the detection of novel *Alpavirinae* sequences. Because these full-length sequences include stretches conserved across proteins from other viral subfamilies, they would probably yield a lower precision. To establish a fair comparison between protein and profile HMM seeds, we assessed the detection ability of GenSeed-HMM using sequences restricted to the VP1R4 seed region (coordinates 799–816—see Supplementary Table 3). The observed individual detection rate was much lower indeed, varying from 0 to 4

TABLE 1 | Taxonomic assignment of contigs (human fecal data, progressive assembly) and classification precision and sensitivity.

Subfamily	Profile HMM seed				
	VP1R1	VP1R4	VP4R1	VP4R3	Total[a]
Alpavirinae	47	43	38	34	65
Gokushovirinae	11	0	11	10	17
Pichovirinae	1	0	1	1	1
Gokush/Alpa[b]	0	0	1	2	2
Total	59	43	51	47	85
VP1R4-positive	40	43	23	24	49
Sensitivity for *Alpavirinae*	(47/65) 72.31%	(43/65) 66.15%	(38/65) 58.46%	(34/65) 52.31%	
Precision for *Alpavirinae*	(47/59) 79.66%	(43/43) 100.00%	(38/51) 74.51%	(34/47) 72.34%	
Sensitivity for VP1R4	(40/49) 81.63%	(43/49) 87.76%	(23/49) 46.94%	(24/49) 48.98%	
Precision for VP1R4	(40/59) 67.80%	(43/43) 100.00%	(23/51) 45.10%	(24/47) 51.06%	

Contigs generated by GenSeed-HMM with the respective profile HMM (VP1R1, VP1R4, VP4R1, and VP4R3) were compared against all reference Microviridae proteins (Roux et al., 2012) using blastx with a cutoff E-value of 1e-20. When hits with similar scores were obtained to more than one subfamily, taxonomic assignment was set to two subfamilies. Contigs were also evaluated for the presence of VP1R4 region by hmmsearch, and the number of positive contigs is shown.

[a] *Total number of De-replicated contigs (See **Figure 3**) that belonged to a given taxonomic assignment.*

[b] *Represents a set of two contigs where the best BLAST hit annotation was below the E-value cutoff and they were equally distant by percent identity to Gokushovirinae and Alpavirinae, so no single assignment was possible.*

sequences with a cutoff of 1e-6, and 0–15 with a cutoff of 1e-2. Although these tests were performed using *blastp* directly instead of running GenSeed-HMM, they show, in a specific manner, that the nature of the seed is what is leading to a difference in sensitivity. These results indicate that a seed-driven assembly based on a single protein sequence is limited to the information contained on that sequence itself, while profile HMMs, by incorporating the variability of a full set or family of sequences, present higher sensitivity and wider range of detection.

Using Multiple Profile HMM Seeds

As presented above, no single profile HMM seed was able to assess the true viral complexity of the sample (**Table 1** and **Figure 3**). Since GenSeed-HMM can use multiple seeds in a single execution, we decided to run a preliminary comparative analysis to evaluate the ability of single and multiple profile HMMs to reconstruct viral genomes. The profile HMMs were employed either individually or in combination of two or four seeds to progressively assemble sequences from the 454 dataset from human fecal samples (Reyes et al., 2010). All identified *Alpavirinae*-specific contigs were submitted to contig size rankings (Supplementary Figure 3), with VP1R1 exhibiting the best overall contig size profile, in agreement with what had been previously observed without filtering out contigs belonging to other *Microviridae* subfamilies (**Figure 2**). When using the VP1R4 and VP4R1 seeds, derived from two distinct viral proteins, the obtained profile was clearly better than the profiles observed with the use of any of the individual seeds. The use of pairs of seeds derived from the same protein (e.g., VP1R1/VP1R4 or VP4R1/VP4R3) did not show relevant improvement over individual seeds (data not shown). Nevertheless, it is noteworthy that a combination of the four seeds yielded the best assembly. This result strongly suggests that a rational combination of profile HMM seeds can be used to unravel the true viral diversity in a sample. Is important to highlight that the number of close to full-length contigs (around 6 kb) does not change with multiple seeds, suggesting that the longest contigs were recovered with either one or multiple seeds.

Phylogenetic Analysis

Using an automated processing pipeline, all sequences assembled with the four profile HMMs were annotated. This automatic annotation was the basis for the identification of the VP1 genes and the respective translation to the corresponding protein sequences. Since we have determined that only the VP1R4 assembly generated sequences restricted to the *Alpavirinae* subfamily (see previous section), this particular annotation set was manually curated and used to produce a dataset of complete and partial VP1 protein sequences. For phylogenetic inference, we used a reference dataset of *Microviridae* proteins (Roux et al., 2012) and sequences publicly available on GenBank (Supplementary Table 1). Since some of the assembled contigs represented incomplete genomes and covered slightly more than the VP1R4 region, we performed two phylogenetic reconstructions using either full-length VP1 proteins or sequences covering approximately 75 amino acids with the VP1R4 region at the C-terminus. The tree containing 28 novel full-length sequences (**Figure 4A**) showed better bootstrap support than that for a tree inferred with 45 shorter sequences (**Figure 4B**), but both converged to the same topology. Both trees clearly separate the different subfamilies and show that all assembled contigs are completely specific to the *Alpavirinae* subfamily, thus corroborating our previous similarity-based taxonomic analysis. In addition, these novel sequences were not

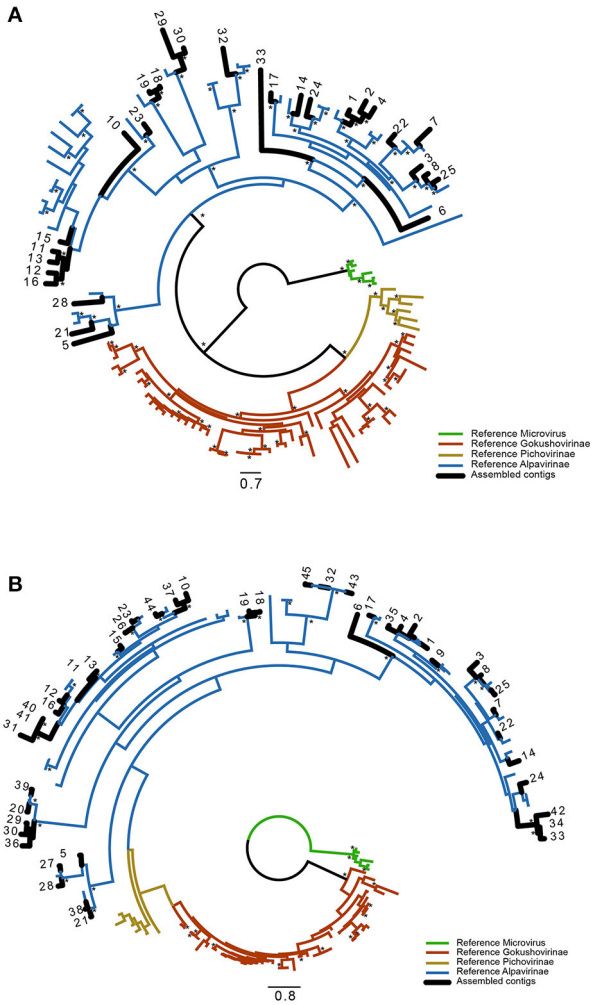


FIGURE 4 | Phylogenetic analysis. Maximum likelihood phylogenetic analysis of **(A)** full-length VP1 protein and **(B)** a shorter region comprising only the VP1R4 HMM region. Sequences were translated from the contigs reconstructed by GenSeed-HMM using the VP1R4 seed. Different subfamilies of the *Microviridae* family are depicted in distinct colors, references were obtained from Roux et al. (2012). Branches represented by sequences derived in this work are labeled in black. Asterisks in the nodes indicate bootstrap values higher than 70%. Numbers represent contig numbers as observed in **Figure 5**.

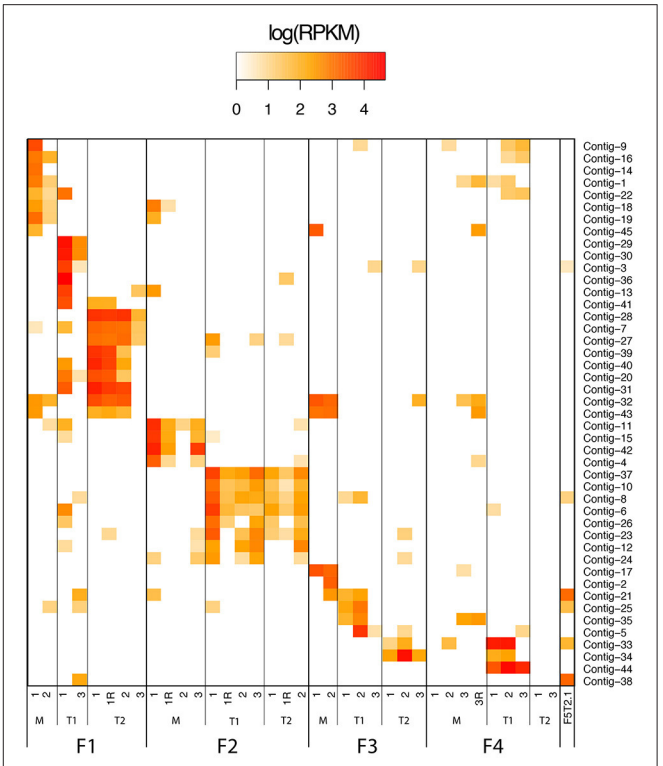


FIGURE 5 | Read abundance of contigs. Heatmap diagram representing read abundance in contigs reconstructed by progressive assembly with GenSeed-HMM and the VP1R4 seed. Fecal biospecimens were collected from different families (F1–F4) composed of monozygotic twins (T1 and T2) and their respective mothers (M). Time points of sample collection and technical replicates (R) are depicted. Data source: 454 dataset from fecal samples of human patients (Reyes et al., 2010).

confined to a few clades, but rather spread in almost all clades containing reference sequences described by Roux et al. (2012) and/or available on GenBank suggesting that this subfamily is highly diverse and broadly dispersed in humans.

Intra- and Inter-Personal Distribution of Novel *Alpavirinae* Sequences

Read abundance in contigs reconstructed by progressive assembly with GenSeed-HMM using the VP1R4 seed was used to assess the distribution of the newly characterized *Alpavirinae* sequences across the different human donor fecal samples. The novel *Alpavirinae* sequences showed highly conserved intrapersonal patterns along different time points (**Figure 5**), similarly to what has been previously observed for whole viromes (Reyes et al., 2010). Conversely, interpersonal viral variations were much higher, with few cases of shared contigs even between twins of the same family, except for the twins on family 2. It has been suggested that the *Alpavirinae* subfamily is linked to genera of the Bacteroidetes phylum (Krupovic and Forterre, 2011). Our results show that distinct individuals harbor different amounts of each of these viruses (**Figure 5**), which are usually not closely phylogenetically related (**Figure 4B**), suggesting that they are probably associated with different Bacteroidetes taxa.

Progressive vs. Global Assembly

To address how progressive assembly performs against conventional global assembly, we compared our contigs generated using GenSeed-HMM with the VP1R4 seed to the results obtained using Newbler in a standalone execution for the same original dataset (global assembly). By selecting only contigs coding for VP1R4-positive proteins (using *hmmsearch*), a fair comparison between both assembly methods could be established. An initial analysis, based on cumulative contig lengths (**Figure 6A**) showed a very similar assembly performance for the 15 longest contigs obtained using the human fecal

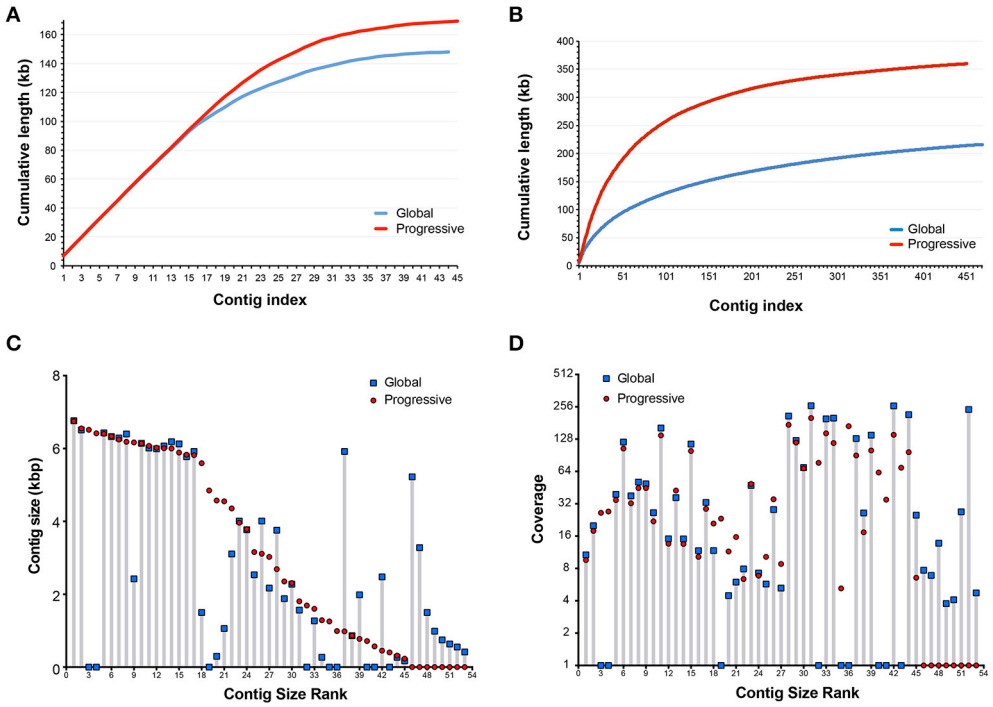


FIGURE 6 | Comparison between global and progressive assembly. Comparison of cumulative contig lengths using progressive assembly with GenSeed-HMM and VP1R4 HMM seed and global assembly with Newbler. Data sources: **(A,C,D)** 454 dataset from fecal samples of human patients (Reyes et al., 2012); **(B)** Illumina dataset from a sewage treatment plant at the municipality of Taboão da Serra, São Paulo, Brazil (unpublished data). Contigs from progressive assembly and VP1R4-positive contigs from global assembly were clustered at 97% identity over at least 90% of the shortest contig, each cluster consisted at most of one contig from each dataset. A total of 53 clusters were generated, nine unique for the progressive assembly and eight unique for global assembly. Plotted is the comparison in lengths **(C)** and coverage **(D)** for related contigs obtained by progressive and global assemblies and ranked by size.

samples. From this result it can be appreciated that the progressive method clearly had a better assembly performance, characterized by a higher number of assembled bases (169 kb) than the global assembly (148 kb). The total number of contigs was similar for both approaches, with 45 contigs in the case of progressive assembly and 44 with global assembly. When the same test was applied to the Illumina dataset derived from a raw sewage sample (**Figure 6B**), we observed even more pronounced differences. In this case, we obtained a total of 360 kb of assembled sequence comprising 453 contigs, whereas the conventional method showed a more fragmented assembly, with a total of 216 kb and 471 contigs (See Supplementary File 2). Given the environmental nature of the raw sewage sample, a much wider viral diversity should be expected. In fact, the total number of contigs was much higher than that observed in human fecal samples. To compare the sensitivity and precision obtained with the GenSeed-HMM method and the global assembly, we performed the same taxonomic annotation and clustering analysis and clustering on these contigs. The results (Supplementary Tables 4, 5) showed equivalent numbers of sensitivity and precision with the VP1R4 seed, confirming that both strategies have similar ability to recover the viral genomes (both are based on the same assembler) but GenSeed-HMM recovers longer contigs with more efficient use of computational resources and completely centered on the target sequences.

To further characterize the consistency among the results obtained with the different strategies, we used the assembly from fecal samples for a similarity clustering at 97% identity to identify cases where the same contig was found in both assemblies. In this case, we observed that each cluster contains at most one contig from each assembly strategy. In the case of the human fecal samples, we identified a total of 53 non-redundant contigs where nine of those were unique to the progressive assembly and eight were unique to the global assembly. When comparing the contig lengths for each pair of clustered contigs (**Figure 6C**) it was possible to see that in 20 cases both assemblies yielded contigs of essentially the same length, while in 11 cases progressive contigs were longer than the global ones, and in five cases the opposite was observed. These findings confirmed once more that the progressive strategy was mostly capable of generating longer contigs from the same original seed than a global strategy. When comparing contig read coverage (**Figure 6D**) in the same dataset, it was clear that both strategies assemble contigs with similar coverage, suggesting that there is no coverage bias for the contigs assembled with the iterative progressive assembly.

DISCUSSION

In this work, we describe the development of GenSeed-HMM, a program that implements a seed-driven progressive assembly

approach using profile HMMs as seeds, in addition to nucleotide and protein sequences. We also demonstrate the application of the implemented method for viral discovery using *Alpavirinae* as a case study. Using a previously published dataset it was possible to assemble a total of 85 *Microviridae* associated contigs, with 25 of those likely representing full viral genomes. Phylogenetic analysis showed that those novel assembled contigs contained representatives of all known clades of the *Alpavirinae* subfamily, significantly contributing to the knowledge regarding those viruses in the human gut. The use of GenSeed-HMM to assemble *de novo* viral genomes present in metagenomes provides a very important resource for the characterization and understanding of the role of different viruses and viral families in the microbial ecology of complex environments.

The current study also shows that our progressive assembly strategy generates an overall higher number of longer contigs, with read coverage equivalent to that observed in the corresponding global assemblies. This improvement could be due to effects of repetitive regions that can create chimeric contigs or even hamper global assembly, especially if these regions are longer than the average read length. Another potential problem is represented by polymorphic sequences, a feature commonly found in viral populations. In the case of global assembly, reads are analyzed all at once to construct the assembly graphs. Conversely, progressive assembly is driven by a single seed or, in the worst case, a relatively small number of seeds. This means that the search space is dramatically reduced since a very strict subset of reads is selected from the main dataset. Each assembled contig then originates two extension seeds, one from each of the contig's ends, which in turn will be used to select new small subsets of reads. Thus, each assembly round employs these relatively few reads plus the previously generated contig, which acts as a guide for sequence growth. Hence, when a repetitive region already present in a previously assembled part of the contig is reached, no newly recruited reads will disrupt the sequence already assembled. The whole process is therefore highly directional, starting from the seed sequence up to the final optimal assembly. This particular *modus operandi* is important to prevent repetitive sequences from leading to chimeric assemblies, which could entrap the process by artificially joining physically unrelated sequences.

A classical protocol for detecting viral sequences from metagenomic data is to assemble the sequence reads and then submit the contigs to BLAST searches against databases of known viral genomes or protein sequences (Cantalupo et al., 2011; Bibby and Peccia, 2013; Norman et al., 2015). This approach is severely limited by the fact that pairwise sequence comparison methods fail to detect distant evolutionary relationships, with sequence identities of around 30% seeming to represent a threshold value for identifying true homologs (Brenner et al., 1998). In the case of viral discovery, this scenario is even more challenging because of the typically high substitution rates, especially in RNA viruses that replicate through error-prone RNA-dependent RNA polymerases (RdRP). Also, the bias of sequence data available for the different viral families limits the effectiveness of similarity searches. Search methods relying on profiles are more sensitive than pairwise alignments because they incorporate broader position-specific information as well as a quantification of the range of substitutions observed across different members of the group. From several methods available, profile HMMs seem to be the most effective to detect distantly related organisms (Park et al., 1998). More recently, Skewes-Cox et al. (2014) reported a method to generate viral profile HMMs (vFams) for the detection of viruses from metagenomic data and the public release of a database composed of more than 4000 such profiles (vFam—http://derisilab.ucsf.edu/software/vFam/). These profile HMMs, constructed from MSAs covering the entire sequence of the respective proteins, showed a higher precision than BLAST searches in real metagenomic datasets, especially for more divergent viral sequences. According to the authors, vFams could be used to nucleate metagenomic assemblies with selected reads to produce longer sequences, in an approach similar to the one previously proposed by our group (Sobreira and Gruber, 2008). Another important aspect pointed out by Skewes-Cox et al. (2014) is the fact that both BLAST and HMM-based methods rely on some degree of similarity to already known viruses, meaning that updating sequence databases in a regular basis is essential for the future effectiveness of such methods, and that bioinformatics approaches based on *de novo* metagenomic assembly and *ab initio* structural prediction algorithms will have increasing importance. In this direction, there is also room to improve seed development with the possible addition of protein structure information in profile HMM design for probing deep phylogenetic associations (Deng and Cheng, 2014).

Compared to the original GenSeed program (Sobreira and Gruber, 2008), the concept of using seeds to drive the assembly process has been extended in GenSeed-HMM by the development of specific routines to deal with profile HMMs. In fact, the originally proposed nucleotide and protein seeds could drive the assembly of sequences derived from the same species or from evolutionarily close organisms. A few previous attempts using our original concepts of seed-driven and/or progressive assembly have been described, but were limited in application to fewer genomic assembly programs, DNA sequence seeds, or non-metagenomic input data (Smits et al., 2015). The original GenSeed program already used both DNA or protein sequences as seeds for iterative assembly, and GenSeed-HMM greatly expands on these capabilities by allowing the use of read data from different sequencing technologies, multiple assemblers, and profile HMM seeds. Tools such as PRICE (Ruby et al., 2013), which also use GenSeed's original assembly principles, are based exclusively on DNA seeds limiting their potential for viral discovery. Indeed, even using protein sequences, which are much more conserved than DNA, the profile HMM seed derived from a short VP1 region (VP1R4) was much more sensitive than any of the 33 complete VP1 protein sequences from Roux et al. (2012) for the detection of novel *Alpavirinae* sequences. Profile HMMs increase the spectrum of detectable organisms since they are built from MSAs derived from many organisms, encompassing a large range of variability within a single probabilistic model. The use of profile HMMs in a targeted gene assembly tool has been recently implemented on the SAT-Assembler program (Zhang et al., 2014). Using a concept similar to the seed-driven assembly described by our

group and implemented in GenSeed (Sobreira and Gruber, 2008), SAT-Assembler uses the seeds to select reads from datasets and then proceeds to construct its own overlap graph for the assembly, also avoiding an all-against-all sequence comparison. However, SAT-Assembler can only generate a consensus sequence that is limited to these reads. Conversely, by means of the progressive assembly method, GenSeed-HMM can extend the sequence reconstruction as much as possible, according to user requirements. This is especially important, since the assembly is not restricted to the gene itself, but also to its flanking regions, providing genomic context information. In fact, by using the appropriate number of assembly cycles, an entire viral (or other episomes, such as mitochondrial) genome can be reconstructed using a single seed, provided that sufficient read coverage is available in the sequencing dataset, as shown in the current study.

When applied to viral discovery, simultaneous use of multiple seeds can substantially increase the sensitivity of the method by generating several starting points for assembly. If maximum sensitivity is required, combining seeds is important, as our results show that no single profile HMM seed can assess the true viral diversity present on any sequencing dataset. However, the proper choice of seeds is essential, since closely placed seeds may be inefficient for two reasons: (1) if the seeds are directed toward physically close regions, chances are that low read coverage may apply to all of them; and (2) because of the physical proximity of the seeds, specific reads recruited by a seed could overlap reads selected by other seeds, implying that the progressive assembly might give rise to something approaching a classical global assembly. Our results show that using seeds derived from different proteins is a more sensible approach. However, it is worth mentioning that using multiple seeds to attain maximum sensitivity may come at the price of lowering precision. A general recommendation for seed design includes avoiding low-complexity regions, as they would result in non-specific reads being recruited and assembled, with a consequent lack of specificity. A good compromise between sequence conservation/divergence of the region selected for profile HMM building may vary from case to case and there is no *a priori* set of rules. Delimiting the range of targeted taxa may help to define whether the profile HMM seeds should be built from selected regions or from a full-length protein sequence. Specific routines could also be implemented in future versions of GenSeed-HMM to identify and discard spurious non-specific sequences. The development of multiple seeds could also profit from a nested, hierarchical-based rationale for seed design and use that should entail aspects of viral taxonomy. For example, one could progressively use sets of seeds, initiating by using replicases, which would then lead to an informed choice of helicase and capsid-derived seeds, and so on. This would drive new virus discovery from core functions, such as replicases and capsid genes (that define viral families) to more contextual functions, such as receptor glycoproteins that would be more informative at the genus level (de Andrade Zanotto and Krakauer, 2008; Krakauer and Zanotto, 2008). We foresee that a rational protocol of profile HMM construction can be established focusing on the development of narrow- and wide-range taxonomic associations. For instance, specific profile HMMs could be built for the detection of well-delimited taxonomic groups such as subfamilies or families.

A paradigm of diagnosis, using either serological or nucleic acid-based methods, is that one can only diagnose organisms that are already known. For instance, given a pathogen to be identified by a serological assay, it is mandatory to first establish which antigens or antibodies will be the targets of detection. Likewise, PCR-based assays rely on previous knowledge of the target sequences to be amplified, and microarray-based assays, such as the Virochip, are based on known hybridization targets. Viruses are biological entities in which evolution can be observed in comparably short spans of time, given their fast rates of mutation and substitution. In fact, since the nineteen-seventies, we have witnessed the emergence of many novel human and animal diseases, such as Acquired Immune Deficiency Syndrome (AIDS) caused by the human immunodeficiency virus (HIV), Ebola virus disease (EVD), among others (Palacios et al., 2008; Wang, 2011; Rosenberg, 2015). Metagenomic data has contributed to surveys of viral diversity (Bibby and Peccia, 2013) and the discovery of novel animal (Belák et al., 2013) and human (Tang and Chiu, 2010; Siebrasse et al., 2012; Phan et al., 2015; Reyes et al., 2015) viruses. The pace of viral discovery is increasing, including many emergent zoonotic viruses pathogenic to humans (Wang, 2011; Rosenberg, 2015). Given the ever-growing amount of sequence data, the challenge is how to diagnose new potentially emerging pathogens without knowing what one is looking for. Considering that emerging viruses moved into humans from pre-existing lineages from the zoonotic pool, some key structures are conserved in essential functions such as replication and capsid proteins. HMMs able to potentially detect a wider range of taxa could be used for epidemiological surveillance, in order to monitor the emergence of new variants of already known viruses or even detect the arising of novel viruses. Profile HMMs have a series of advantages that make them ideally suited to detect sequences that have not been sampled in the original MSA within a reasonable margin of divergence, detecting related members to those used for the construction profile that likely share the same selective pressures. This feature opens up a new possibility, namely the diagnosis of novel viruses potentially pathogenic to humans and animals, without previous specific information, an approach that we refer to as *de novo* diagnosis. We believe that *de novo* diagnosis using rationally designed profile HMMs may assume a fundamental importance for epidemiological surveillance in some sentinel sites such as hospitals, sewage treatment stations, animal production facilities, and migratory bird colonies, among others. By detecting emerging viruses on these sites, it would be possible to undertake containment measures to prevent the spread of potentially devastating diseases. GenSeed-HMM provides a fast and simple implementation to run progressive assembly pipelines using profile HMMs covering the most relevant groups of viral pathogens. By combining rational design of profile HMMs and multiple GenSeed-HMM runs, one can foresee a replacement of the paradigm of conventional diagnosis.

In this work we exemplified how GenSeed-HMM could be used for viral discovery. Nonetheless, the spectrum of potential

applications of the seed-driven progressive assembly method using profile HMMs is much wider. Besides viral genomes, the method is well fitted for surveys of extra-chromosomal elements such as plastid and mitochondrial genomes from metagenomic data. This is particularly relevant for the exploration of some specific target sequences from largely contaminated datasets such as paleometagenomic samples. Profile HMM seeds can also be used to reconstruct specific protein coding genes for gene diversity studies, thus determining all possible gene variants present in a metagenomic sample, independently of their organism of origin. Such surveys could be useful to detect the emergence of drug-resistant variants in sensitive environments such as hospitals and animal production facilities, where antibiotics are regularly used. In addition, the extra length obtained with iterative progressive assembly of these target-specific sequences could reveal their genomic context, that is, whether they are originated from chromosomal or episomal sources, and surrounded by other genes involved with drug-resistance and/or associated with transposable elements. By using multiple profile HMM seeds, built from proteins from a specific pathway, GenSeed-HMM allows one to assess the occurrence of this pathway in specific environmental metagenomic samples, even if the gene complement is derived from multiple organism sources. Finally, another interesting application is the use of the progressive assembly method as an adjunct for gap closure on assembly finishing projects, by using multiple contig ends as anchored seeds to promote a sequence walking/progressive assembly process in which overlapping sequences can lead to gap closure. Using an in-house script for this specific application, we were able to close around 80% of the gaps of a bacterial sequencing project (data not shown). Concluding, GenSeed-HMM is a multipurpose program under active development, and we envisage its growing application on a variety of forthcoming projects.

AUTHOR CONTRIBUTIONS

AG and AR conceived and designed the study. AMD, JA, and PZ contributed to the design of the experiments. AG, AR, ALO, JMG, JA, LO, MT, and TS performed the experiments. AG, AR, ALO, JMG, JA, and TS analyzed the data. DM and EM collected raw sewage samples and generated sequencing data. AG and AR prepared the first draft of the manuscript. JA and PZ participated in the discussion and writing of the manuscript. All authors revised the manuscript and have agreed to the final content.

ACKNOWLEDGMENTS

AG, PZ, and AMD received Productivity-in-Research fellowships from the National Council for Scientific and Technological Development (CNPq). TS received an IC scholarship from PIBIC/CNPq. EM received a DT scholarship from CAPES. JA is supported by grant #2013/14622-3, São Paulo Research Foundation (FAPESP). AR is supported by FAPA internal funding at Universidad de los Andes. JMG is supported by Young Investigator award from Colciencias and the School of Sciences at Universidad de los Andes. ALO received an MS scholarship from CAPES and São Paulo Research Foundation - FAPESP (#2010/04609-1).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2016.00269

REFERENCES

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Belák, S., Karlsson, O. E., Blomström, A. L., Berg, M., and Granberg, F. (2013). New viruses in veterinary medicine, detected by metagenomic approaches. *Vet. Microbiol.* 165, 95–101. doi: 10.1016/j.vetmic.2013.01.022

Bexfield, N., and Kellam, P. (2011). Metagenomics and the molecular identification of novel viruses. *Vet. J.* 190, 191–198. doi: 10.1016/j.tvjl.2010.10.014

Bibby, K., and Peccia, J. (2013). Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environ. Sci. Technol.* 47, 1945–1951. doi: 10.1021/es305181x

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U.S.A.* 95, 6073–6078. doi: 10.1073/pnas.95.11.6073

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

Cantalupo, P. G., Calgua, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., et al. (2011). Raw sewage harbors diverse viral populations. *mBio* 2:e00180-11. doi: 10.1128/mBio.00180-11

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. doi: 10.1093/bioinformatics/btr088

de Andrade Zanotto, P. M., and Krakauer, D. C. (2008). Complete genome viral phylogenies suggests the concerted evolution of regulatory cores and accessory satellites. *PLoS ONE* 3:e3500. doi: 10.1371/journal.pone.0003500

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* 23, 673–679. doi: 10.1093/bioinformatics/btm009

Deng, X., and Cheng, J. (2014). Enhancing HMM-based protein profile-profile alignment with structural features and evolutionary coupling information. *BMC Bioinformatics* 15:252. doi: 10.1186/1471-2105-15-252

Durham, A. M., Kashiwabara, A. Y., Matsunaga, F. T., Ahagon, P. H., Rainone, F., Varuzza, L., et al. (2005). EGene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics* 21, 2812–2813. doi: 10.1093/bioinformatics/bti424

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 4498. doi: 10.1038/ncomms5498

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

El-Metwally, S., Hamza, T., Zakaria, M., and Helmy, M. (2013). Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput. Biol.* 9:e1003345. doi: 10.1371/journal.pcbi.1003345

Fancello, L., Raoult, D., and Desnues, C. (2012). Computational tools for viral metagenomics and their application in clinical research. *Virology* 434, 162–174. doi: 10.1016/j.virol.2012.09.025

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Kim, M. S., Park, E. J., Roh, S. W., and Bae, J. W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* 77, 8062–8070. doi: 10.1128/AEM.06331-11

Krakauer, D. C., and Zanotto, P. (2008). "Viral individuality and limitations of the life concept," in *Protocells: Bridging Nonliving and Living Matter*, eds S. Rasmussen, M. A. Bedau, L. Chen, D. Deamer, D. C. Krakauer, N. H. Packard, and P. F. Stadler (Cambridge, MA: MIT Press Scholarship Online), 513–536.

Krupovic, M., and Forterre, P. (2011). Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS ONE* 6:e19893. doi: 10.1371/journal.pone.0019893

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lipkin, W. I., and Firth, C. (2013). Viral surveillance and discovery. *Curr. Opin. Virol.* 3, 199–204. doi: 10.1016/j.coviro.2013.03.010

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* 1:18. doi: 10.1186/2047-217X-1-18

Mayer, E. A., Knight, R., Mazmanian, S. K., Cryan, J. F., and Tillisch, K. (2014). Gut microbes and the brain: paradigm shift in neuroscience. *J. Neurosci.* 34, 15490–15496. doi: 10.1523/JNEUROSCI.3299-14.2014

Mehnert, D. U., and Stewien, K. E. (1993). Detection and distribution of rotavirus in raw sewage and creeks in Sao Paulo, Brazil. *Appl. Environ. Microbiol.* 59, 140–143.

Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., et al. (2013). Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinformatics* 14, 193–202. doi: 10.1093/bib/bbs012

Mitchell, A., Chang, H. Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* 43, D213–D221. doi: 10.1093/nar/gku1243

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., et al. (2015). Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* 160, 447–460. doi: 10.1016/j.cell.2015.01.002

Palacios, G., Druce, J., Du, L., Tran, T., Birch, C., Briese, T., et al. (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *N. Engl. J. Med.* 358, 991–998. doi: 10.1056/NEJMoa073785

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., et al. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* 284, 1201–1210. doi: 10.1006/jmbi.1998.2221

Phan, T. G., Mori, D., Deng, X., Rajindrajith, S., Ranawaka, U., Fan Ng, T. F., et al. (2015). Small circular single stranded DNA viral genomes in unexplained cases of human encephalitis, diarrhea, and in untreated sewage. *Virology* 482, 98–104. doi: 10.1016/j.virol.2015.03.011

Quaiser, A., Dufresne, A., Ballaud, F., Roux, S., Zivanovic, Y., Colombet, J., et al. (2015). Diversity and comparative genomics of Microviridae in Sphagnum-dominated peatlands. *Front. Microbiol.* 6:375. doi: 10.3389/fmicb.2015.00375

Queiroz, A. P., Santos, F. M., Sassaroli, A., Hársi, C. M., Monezi, T. A., and Mehnert, D. U. (2001). Electropositive filter membrane as an alternative for the elimination of PCR inhibitors from sewage and water samples. *Appl. Environ. Microbiol.* 67, 4614–4618. doi: 10.1128/AEM.67.10.4614-4618.2001

Reyes, A., Blanton, L. V., Cao, S., Zhao, G., Manary, M., Trehan, I., et al. (2015). Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci. U.S.A.* 112, 11941–11946. doi: 10.1073/pnas.1514285112

Reyes, A., Haynes, M., Hanson, N., Angly, F. E., Heath, A. C., Rohwer, F., et al. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466, 334–338. doi: 10.1038/nature09199

Reyes, A., Semenkovich, N. P., Whiteson, K., Rohwer, F., and Gordon, J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nat. Rev. Microbiol.* 10, 607–617. doi: 10.1038/nrmicro2853

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

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

Rosenberg, R. (2015). Detecting the emergence of novel, zoonotic viruses pathogenic to humans. *Cell. Mol. Life Sci.* 72, 1115–1125. doi: 10.1007/s00018-014-1785-y

Roux, S., Krupovic, M., Poulet, A., Debroas, D., and Enault, F. (2012). Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE* 7:e40418. doi: 10.1371/journal.pone.0040418

Ruby, J. G., Bellare, P., and Derisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3 (Bethesda)* 3, 865–880. doi: 10.1534/g3.113.005967

Siebrasse, E. A., Reyes, A., Lim, E. S., Zhao, G., Mkakosya, R. S., Manary, M. J., et al. (2012). Identification of MW polyomavirus, a novel polyomavirus in human stool. *J. Virol.* 86, 10321–10326. doi: 10.1128/JVI.01210-12

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res.* 19, 1117–1123. doi: 10.1101/gr.089532.108

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., and DeRisi, J. L. (2014). Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* 9:e105067. doi: 10.1371/journal.pone.0105067

Smits, S. L., Bodewes, R., Ruiz-González, A., Baumgärtner, W., Koopmans, M. P., Osterhaus, A. D., et al. (2015). Recovering full-length viral genomes from metagenomes. *Front. Microbiol.* 6:1069. doi: 10.3389/fmicb.2015.01069

Smits, S. L., and Osterhaus, A. D. (2013). Virus discovery: one step beyond. *Curr. Opin. Virol.* 3, e1–e6. doi: 10.1016/j.coviro.2013.03.007

Sobreira, T. J., and Gruber, A. (2008). Sequence-specific reconstruction from fragmentary databases using seed sequences: implementation and validation on SAGE, proteome and generic sequencing data. *Bioinformatics* 24, 1676–1680. doi: 10.1093/bioinformatics/btn283

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690. doi: 10.1093/bioinformatics/btl446

Tang, P., and Chiu, C. (2010). Metagenomics for the discovery of novel human viruses. *Future Microbiol.* 5, 177–189. doi: 10.2217/fmb.09.120

Wajid, B., and Serpedin, E. (2012). Review of general algorithmic features for genome assemblers for next generation sequencers. *Genomics Proteomics Bioinformatics* 10, 58–73. doi: 10.1016/j.gpb.2012.05.006

Wang, L. F. (2011). Discovering novel zoonotic viruses. *N. S. W. Public Health Bull.* 22, 113–117. doi: 10.1071/NB10078

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, Y., Sun, Y., and Cole, J. R. (2014). A scalable and accurate targeted gene assembly tool (SAT-Assembler) for next-generation sequencing data. *PLoS Comput. Biol.* 10:e1003737. doi: 10.1371/journal.pcbi.1003737

Zhong, X., Guidoni, B., Jacas, L., and Jacquet, S. (2015). Structure and diversity of ssDNA Microviridae viruses in two peri-alpine lakes (Annecy and Bourget, France). *Res. Microbiol.* 166, 644–654. doi: 10.1016/j.resmic.2015.07.003

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

frontiers in MICROBIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 December 2014
doi: 10.3389/fmicb.2014.00714

Assembly of viral genomes from metagenomes

Saskia L. Smits[1,2†], Rogier Bodewes[1†], Aritz Ruiz-Gonzalez[3,4,5], Wolfgang Baumgärtner[6], Marion P. Koopmans[1,7], Albert D. M. E. Osterhaus[1,2,8] and Anita C. Schürch[1]*

[1] Department of Viroscience, Erasmus Medical Center, Rotterdam, Netherlands
[2] Viroclinics Biosciences, Rotterdam, Netherlands
[3] Department of Zoology and Animal Cell Biology, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain
[4] Systematics, Biogeography and Population Dynamics Research Group, Lascaray Research Center, University of the Basque Country (UPV/EHU), Vitoria-Gasteiz, Spain
[5] Conservation Genetics Laboratory, National Institute for Environmental Protection and Research (ISPRA), Bologna, Italy
[6] Department of Pathology, University of Veterinary Medicine Hannover, Hannover, Germany
[7] Centre for Infectious Diseases Research, Diagnostics and Screening, National Institute for Public Health and the Environment, Bilthoven, Netherlands
[8] Center for Infection Medicine and Zoonoses Research, Hannover, Germany

Edited by:
Richard J. Hall, Institute of Environmental Science and Research, New Zealand

Reviewed by:
Hirokazu Kimura, National Institute of Infectious Diseases, Japan
Karen Dawn Weynberg, Australian Institute of Marine Science, Australia
Patrick Jon Biggs, Massey University, New Zealand

***Correspondence:**
Anita C. Schürch, Department of Viroscience, Erasmus Medical Center, PO Box 2040, Rotterdam 3000 CA, Netherlands
e-mail: a.schurch@erasmusmc.nl

† *These authors have contributed equally to this work.*

Viral infections remain a serious global health issue. Metagenomic approaches are increasingly used in the detection of novel viral pathogens but also to generate complete genomes of uncultivated viruses. *In silico* identification of complete viral genomes from sequence data would allow rapid phylogenetic characterization of these new viruses. Often, however, complete viral genomes are not recovered, but rather several distinct contigs derived from a single entity are, some of which have no sequence homology to any known proteins. *De novo* assembly of single viruses from a metagenome is challenging, not only because of the lack of a reference genome, but also because of intrapopulation variation and uneven or insufficient coverage. Here we explored different assembly algorithms, remote homology searches, genome-specific sequence motifs, k-mer frequency ranking, and coverage profile binning to detect and obtain viral target genomes from metagenomes. All methods were tested on 454-generated sequencing datasets containing three recently described RNA viruses with a relatively large genome which were divergent to previously known viruses from the viral families *Rhabdoviridae* and *Coronaviridae*. Depending on specific characteristics of the target virus and the metagenomic community, different assembly and *in silico* gap closure strategies were successful in obtaining near complete viral genomes.

Keywords: virus, pathogen, metagenome, virome, virus discovery, assembly, viral metagenomics

INTRODUCTION

Human and animal populations are continuously confronted with emerging viral infections (Delwart, 2007; Lipkin, 2010; Smits and Osterhaus, 2013). In a proportion of patients and animals suffering from disease, no pathogens can be detected using a range of sensitive diagnostic assays, suggesting the presence of unidentified viruses in human and animal populations (Bloch and Glaser, 2007; Denno et al., 2012). Classically, new viruses were identified by standard molecular detection methods, virus replication in tissue culture or animal experiments. Nowadays, in order to discover and characterize new or (re-) emerging viruses, metagenome sequencing is increasingly being used to identify viral pathogens. In addition, these techniques are more and more often being used to generate complete genomes of uncultivated viruses, but also other organisms (Delwart, 2007; Lipkin, 2010; Iverson et al., 2012; Albertsen et al., 2013; Smits and Osterhaus, 2013; Handley et al., 2014).

Metagenomic strategies to virus discovery rely on sequence-independent amplification of nucleic acids combined with next generation sequencing platforms instead of targeting specific genomic loci, thereby generating DNA sequences (i.e., reads) that align to various genomic locations for the numerous genomes present in the sample, including non-microbes (Sharpton, 2014). Common random amplification methods are multiple displacement amplification (MDA) or sequence-independent single-primer amplification (SISPA) (Hutchison et al., 2005; Spits et al., 2006; Delwart, 2007; Djikeng et al., 2008; Lipkin, 2010; Smits and Osterhaus, 2013). The advantages of sequence-independent amplification are simplicity and relative speed and the ability to identify and sequence hundreds of viruses simultaneously thereby allowing detection of new or previously unrecognized viruses that are highly divergent from already described ones (Bodewes et al., 2014a,c). Inherent to the approach is that a large fraction of the metagenome consists of sequences of other organisms than the viral targets, including host sequences, archaea, bacteria, and bacteriophages, despite physical enrichment strategies for virus particles that are often applied (Van Leeuwen et al., 2010; Kostic et al., 2012; Van Den Brand et al., 2012; Wylie et al., 2012; Bodewes et al., 2013; Schurch et al., 2014).

Metagenomic sequence data analysis with the aim to identify viral sequences presents several challenges. Datasets are relatively complex and large. In addition, the obtained viral

reads in metagenomes can either originate from taxonomically informative genomic regions and even provide insight in the biological function of the encoded protein or originate from less conserved genomic regions for which biological functions are difficult to assign. Current strategies rely mostly on filtering steps to remove host nucleic acid from metagenomes either before or after sequencing and analysis of the data, including assembly and homology searches against annotated sequences in public databases (Woyke et al., 2006; Chew and Holmes, 2009; Schmieder and Edwards, 2012; Garcia-Garcera et al., 2013; Prachayangprecha et al., 2014; Schurch et al., 2014). Untargeted metagenomic approaches have enabled the identification of numerous newly emerging or previously unidentified viral pathogens in recent years. However, obtaining full-length viral genomes from metagenomic datasets remains challenging.

The number of reads obtained from a specific virus in metagenome samples is correlated to the viral load in the sample under study (De Vries et al., 2012; Prachayangprecha et al., 2014). In some cases, the number of reads in the sample is sufficient to permit enough read overlaps to establish longer contiguous sequences (contigs). However, direct assembly of complete viral and bacterial genomes from metagenomic data can involve a large amount of manual curation (Handley et al., 2014; Sharpton, 2014) as most pathogen genomes are not completely represented by reads and most viral communities are highly diverse (Mavromatis et al., 2007; Mende et al., 2012). Currently, full-length viral genomes are often obtained with additional experimental approaches based on PCR amplification with specific primers designed on obtained reads or contigs and/or 5′ and 3′ RACE PCR in combination with a Sanger sequencing approach (Van Leeuwen et al., 2010; Siegers et al., 2014). However, by optimally mining sequences in metagenomes, the likelihood and speed of identifying viral reads and the level of viral genome completeness can be increased and the need for laboratory follow-up minimized. In the present study, we describe and compare methods to obtain viral target genomes from metagenomes using a retrospective approach on 454-sequencing datasets containing three recently described viruses from the families *Rhabdo-* and *Coronaviridae*.

METHODS

DATASETS

The first metagenome dataset was obtained from a cell culture supernatant (CCS) containing a rhabdovirus-like virus isolated from tissue collected from a stranded white-beaked dolphin (*Lagenorhynchus albirostris*) (Osterhaus et al., 1993; Siegers et al., 2014). Genetic and phylogenetic characterization of the dolphin rhabdovirus (DRV) revealed that it was closely related to rhabdoviruses of the genera *Perhabdovirus* and *Vesiculovirus* found in fish (Siegers et al., 2014). In the second case, a highly divergent rhabdovirus, called red fox fecal rhabdovirus (RFFRV) was identified during a metagenomic survey of feces of red foxes from Spain (*Vulpes vulpes*) (Bodewes et al., 2014c). The last metagenome dataset was from lung tissue of a dead Indian python (*Python molurus*) with pneumonia, in which a novel nidovirus belonging to the family *Coronaviridae* within the order *Nidovirales* was identified. It was the first description of a reptile nidovirus (python nidovirus, PNV) and phylogenetic analysis placed this virus in the subfamily *Torovirinae* (Bodewes et al., 2014a). These datasets were acquired using a random sequence amplification and deep sequencing approach on a 454 GS Junior instrument (Roche) as previously described by Van Leeuwen et al. (2010), Bodewes et al. (2013, 2014a,c). At present full-length genomes (DRV) or expected complete coding sequences (PNV, RFFRV) are available.

ASSEMBLY METHODS

Four different assembly methods, exhaustive iterative assembly (Schurch et al., 2014), CLC Genomics Workbench 6.0.4 assembler (CLC bio, Aarhus, Denmark), Genovo version 0.4 (Laserson et al., 2011), and Newbler 2.5 (Roche), were compared in their efficiency of detecting viral reads in the three metagenome datasets. The originally used method was iterative exhaustive assembly. Iterative exhaustive assembly of sequences is part of a virus discovery pipeline written in the python programming language (Python 2.7) that includes trimming of reads and initial assembly with Newbler (454GS Assembler version 2.7, Roche), with standard parameters. Trimmed reads and initial contigs were subjected to assembly by CAP3 (VersionDate: 12/21/07) (Huang and Madan, 1999) with standard parameters. The resulting singletons and contigs were iteratively assembled by CAP3 until no new contigs were formed.

Subsequently, the trimmed reads were mapped back to the identified taxonomic units with Newbler (454 GSMapper version 2.7, Roche) with standard parameters (Schurch et al., 2014). CLC Genomics Workbench 6.0.4 assembler (CLC bio, Aarhus, Denmark) was run with the previously trimmed reads with automatic bubble and word size. Genovo version 0.4 was run with 40 iterations and otherwise default values (Laserson et al., 2011). Newbler 2.5 (Roche) was run with default values.

DETERMINATION OF TAXONOMIC CONTENT

Contigs and singletons of the iterative assembly approach that were longer than 75 bases were filtered with Dustmasker which is part of the NCBI-BLAST+ 2.2.25 suite of tools for sequences that contain more than 60% low complexity sequences (Camacho et al., 2009). After filtering of low complexity sequences, the remaining taxonomic units were subjected to a BLASTN search against a database that contained only nucleotide sequences from birds (Aves, taxonomic identifier 8782), carnivores (Carnivora, taxID 33554), primates (Primates, taxID 9443), rodents (Rodentia, taxID 9989), and ruminants (Ruminantia, taxID 9845) with an *e*-value cut-off of 0.001 for subtraction of potential host sequences. Sequences without hits in the host-BLAST were then subjected to a BLASTN search against the entire nt database with an *e*-value cut-off of 0.001. All sequences without hits were then subjected to a BLASTX search against protein sequences present in the nr database. BLAST hits were categorized by assigning a taxonomic category.

The percentage of viral reads in the sequence datasets and read coverage of the target genome using different assembly methods were determined by mapping trimmed reads to reference

genomes with GSMapper Version 2.7 (Roche) with a minimum overlap identity of 80%.

REMOTE HOMOLOGY SEARCH

All contigs were translated in six frames. Hidden Markov Models (HMMs) of PFAM families associated with *Rhabdoviridae* (pfam14314, pfam00945, pfam02484, pfam03216, pfam03342, pfam03012, pfam03397, pfam04785, pfam05554, pfam00922, pfam00974, pfam06326) were used to search the translated contigs of the metagenome datasets with rhabdoviruses with HMMER3.1 (Punta et al., 2012). Accordingly, HMMs of 45 PFAM families associated with *Coronaviridae* (pfam05213, pfam06460, pfam04694, pfam09408, pfam08717, pfam08716, pfam08715, pfam06478, pfam06471, pfam05409, pfam03262, pfam03053, pfam02723, pfam01601, pfam01600, pfam00937, pfam08779, pfam12383, pfam12379, pfam12133, pfam12124, pfam12093, pfam11963, pfam11633, pfam11501, pfam11395, pfam11289, pfam11030, pfam10943, pfam09401, pfam08710, pfam06336, pfam06145, pfam05528, pfam04753, pfam03905, pfam03622, pfam03620, pfam03617, pfam03187, pfam02398, pfam01635, pfam09399, pfam01831) were used to search the translated contigs of the PNV metagenome.

MOTIF DISCOVERY AND MOTIF SEARCH

Motif sequence patterns were discovered with MEME Version 4.9.1 (Bailey et al., 2009) by allowing any number of repetitions on the sequence. The best scoring detected motif distributed over the seed contig was then used to search the motif in the collection of all contigs longer than 500 bases in all three datasets with MAST (Bailey et al., 2009) with an *e*-value lower than 0.1.

COVERAGE PROFILE BINNING

The average coverage of all contigs identified using exhaustive iterative assembly was calculated by dividing the number of reads covering the contigs by its length, as determined by the mapping procedure of the virus discovery pipeline. Frequency of binned coverage profiles was visualized in R statistical package version 3.1.

K-mer FREQUENCY RANKING

K-mer frequency was determined with the Bioconductor package biostrings (Pages et al., in press) for 3mers to 8mers for contigs larger than 1 kb in R statistical package version 3.1 (Team, 2012). Absolute differences between the k-mer frequencies of the seed contig and all other contigs were summed among different k-mer lengths and ranked, and visualized in relation to contig size.

ACCESSION NUMBERS

Viral genome sequences used in this study were taken from Genbank, accession numbers KF958252 (DRV), KF823814 (RFFRV), and KJ935003 (PNV).

RESULTS

EVALUATION OF DIFFERENT ASSEMBLY ALGORITHMS

The objective of this study was to test and evaluate methods to increase the likelihood and speed of identifying viral reads and the level of viral genome completeness from metagenomic datasets generated on the 454-sequencing platform. The three 454-sequencing datasets obtained from a CCS, a red fox fecal (RFF) metagenome, and a tiger python lung tissue (TPLT) metagenome contained 69,358, 56,174, and 135,812 sequence reads, respectively (**Table 1**). These reads were analyzed with an automatic analysis pipeline that included stringent quality and length trimming, exhaustive iterative assembly, and low complexity filtering (Schurch et al., 2014). A total of 28,207, 32,455 and 50,024 reads from the CCS and the RFF and TPLT metagenome, respectively were subjected to homology searches (**Table 2**). The analysis showed a high variety among almost all taxonomic categories in the three different datasets (**Table 2**). The overall viral content determined by homology search was relatively low (0.72%) in dataset 3 (TPLT metagenome), and high (30.21 and 68.05%) in datasets 1 and 2 (CCS and RFF metagenome, **Table 2**).

Iterative exhaustive assembly resulted in assembled metagenomes containing between 40–60% of the original total amount of obtained reads (**Table 1**) of which the virome has a large dynamic range of reads depending on the sample under analysis. Unsurprisingly, the CCS dataset from an assumingly relatively pure virus culture supernatant had a high viral content, consisting predominantly of DRV. The viromes of the RFF metagenome showed a much smaller percentage of the RFFRV indicating the presence of multiple different viruses (**Tables 1**, **2**).

Genomes of DRV, RFFRV, and PNV were not completely assembled by the exhaustive iterative assembly approach implemented in the automated analysis pipeline. The largest contigs

Table 1 | Description of deep sequencing datasets.

	CCS	RFF	TPLT
Total number of reads	69358	56174	135812
Assembled metagenome (%)	40.67	57.78	36.83
Reads identified by homology search as obtained from target virus (%)	27.67	5.82	0.11
Reads retrospectively obtained from target virus (%)	69.52	13.58	26.14

Cell culture supernatant (CCS) containing Dolphin rhabdovirus (DRV), red fox feces (RFF) metagenome containing red fox fecal rhabdovirus (RFFRV) and python lung tissue (TPLT) metagenome containing python nidovirus (PNV).

Table 2 | Taxonomic composition of deep sequencing datasets.

	CCS	RFF	TPLT
Unassigned	0.04	0.19	0.97
Virus	68.05	30.21	0.72
Unknown	3.90	10.35	49.39
Eukaryota	27.40	37.90	35.22
Bacteria	0.61	21.34	13.64
Archea	0	0	0.06

Taxonomic composition per read in percentage of cell culture supernatant (CCS) containing Dolphin rhabdovirus (DRV), red fox feces (RFF) metagenome containing red fox fecal rhabdovirus (RFFRV) and python lung tissue (TPLT) metagenome containing python nidovirus (PNV). Unassigned, best BLAST hit without taxonomic assignment. Unknown, no homology to any database entry.

were 7291 bases (64.32% of DRV, DRV seed contig) and 7682 bases (47.9% of RFFRV), respectively, of an expected size of 11 to 15 kb for *Rhabdoviridae* and 24,734 bases (73.68% of PNV) of an expected 30 kb for *Coronaviridae* (**Figures 1A–C**). Interestingly, retrospective mapping of reads to the viral target genomes showed that a large percentage of the sequences identified as "unknowns" by homology searches in the TPLT and RFF metagenome were actually derived from the target genome, most likely from parts of the target genomes without detectable similarity to any other viral protein in the BLAST database (**Tables 1**, **2**).

To evaluate if other assembly algorithms would be able to directly assemble the complete target viral genomes from the deep sequencing data, we compared the contigs assembled from trimmed reads by iterative assembly, CLC Genomics Workbench assembler, Genovo and Newbler. While Genovo and Newbler both produced many small contigs covering part of the target genomes (**Figures 1A–C**), CLC Genomics Workbench assembler and the iterative assembly approach produced a similar number of contigs (two to five). However, large contigs (>1 kb) produced by iterative assembly covered the target genomes more completely than any other set of contigs obtained with other assembly algorithms. None of the assemblers tested here was able to completely assemble the viral genomes from the reads into a single contig (**Figures 1A–C**).

The contigs produced by iterative assembly and CLC Genomics Workbench for DRV were clearly overlapping (**Figure 1A**) and could be fused to a single assembly of a complete DRV genome by manual curation. For RFFRV in dataset 2, contigs assembled by iterative assembly and Genovo overlapped. However, a very small overlap of only five nucleotides between position 4356 and 4361, probably due to the combination of a drop in coverage and a stretch of sequence with low complexity (**Figure 1B**) did not allow us to retrieve a complete viral genome. Moreover, with the exception of the largest contig (the RFFRV seed contig), no other RFFRV contigs had a homolog in the NCBI nucleotide or protein database. The minor overlap, in combination with absence of homology, prevented assembly of a complete RFFRV genome. Similarly, the overlaps between contigs of PNV obtained with different assembly algorithms, in combination with the absence of homology, were insufficient to conclusively obtain a full-length PNV genome. In the TPLT metagenome, a 24.7 kb contig had a stretch with an average identity of 29% amino acid identity to the replicase polyprotein of Berne virus, subfamily *Torovirinae* (**Figure 1C**). This contig was used as seed contig. Overall, the data indicate that iterative exhaustive assembly seems to perform best in terms of production of large contigs and coverage of target genomes compared to other assemblers. Thus, for further analysis we used the set of contigs produced by iterative assembly. It is of note, however, that using a combination of different assembly algorithms may result in a higher level of completeness of target genomes if not complete genome assembly.

REMOTE HOMOLOGY SEARCH

In absence of BLAST-detectable sequence homology to previously described viruses for some stretches of the target genomes we attempted to use methods that are able to detect remote homologs, i.e., profile hidden markov models (**Figures 1A–C**). To retrieve and link contigs of the viral target genomes we used profile HMMs of protein domains present in *Rhabdoviridae* and *Coronaviridae*, respectively. For *Rhabdoviridae*, 12 domains were present in PFAM, from nucleocapsid, spike and matrix proteins. Searching the translated contigs of the CCS containing DRV identified three regions (Vesiculovirus matrix protein—PF06326.7, Rhabdovirus nucleocapsid protein—PF00945.13 and the Rhabdovirus spike protein—PF00974.13) covered by several contigs (**Figure 1A**, contigs smaller than 1 kb not shown). The RFF metagenome did not give any hits (**Figure 1B**). For *Coronaviridae* in the TPLT metagenome, 44 HMMs were present in PFAM, again covering all proteins families. Three *Coronaviridae*-specific HMMs identified several contigs, the Coronavirus NSP13 (F06460.7), the RNA synthesis protein NSP10 (PF09401.5) and the Coronavirus RPol N-terminus (PF06478.8). However, in all three datasets, all identified domains were already identified in contigs with BLAST sequence homology to a closely related virus (**Figures 1A–C**). Identification of regions with remote homology to family-specific domains did therefore not result in acquisition of additional genomic regions that were not identified by the original iterative assembly method in combination with homology search by BLAST.

MOTIF SEARCH

A sequence motif is a DNA pattern that occurs repeatedly in a genome or in a group of related sequences. *De novo* motif discovery is independent of previously described motifs and their function. Motif discovery was performed on the seed contigs that showed homology to viruses of either the *Rhabdoviridae* or *Coronaviridae* family of metagenome 1 and 3 and on two adjacent, clearly overlapping RFFRV contigs of metagenome 2, including the RFFRV seed contig (**Figure 2B**). The highest scoring motif (**Figures S1A–C**) of each seed contig was then used to screen all available contigs of the deep sequencing datasets. Contigs were selected if they contained one or several occurrences of the motif at an *e*-value smaller than 0.01. Four additional DRV-matching contigs smaller than 1 kb (not indicated in **Figure 1A**) were identified in the CCS. In the RFF metagenome, four additional RFFRV contigs were identified and in the TLPT metagenome, one additional PNV contig exhibited the detected pattern (**Figures 1B–C**). No false positive contigs were identified with this method. Moreover, when the PNV motif was used to search all three viral genomes including the rhabdovirus genomes, it exhibited highest specificity for the PNV genome (*e*-value 4e-12), with *e*-values above 1 for RFFRV and DRV. The DRV and the RFFRV motif were most specific for their originating genome, while the *e*-value for the respective other rhabdovirus genome was relatively low (0.62 and 0.015), suggesting those motifs might be conserved among both rhabdovirus genomes. The high specificity of the motifs described here was also demonstrated when scanning all contigs of the three metagenomes: the motif discovered contigs of the respective viral genomes with a high specificity. The detection sensitivity however was restricted by the number of contigs that contained the motif. For example, the eight occurrences of the motif in the RFFRV genome were, with one exception, all found in intergenic regions. Contigs not containing intergenic regions could not be identified with this method.

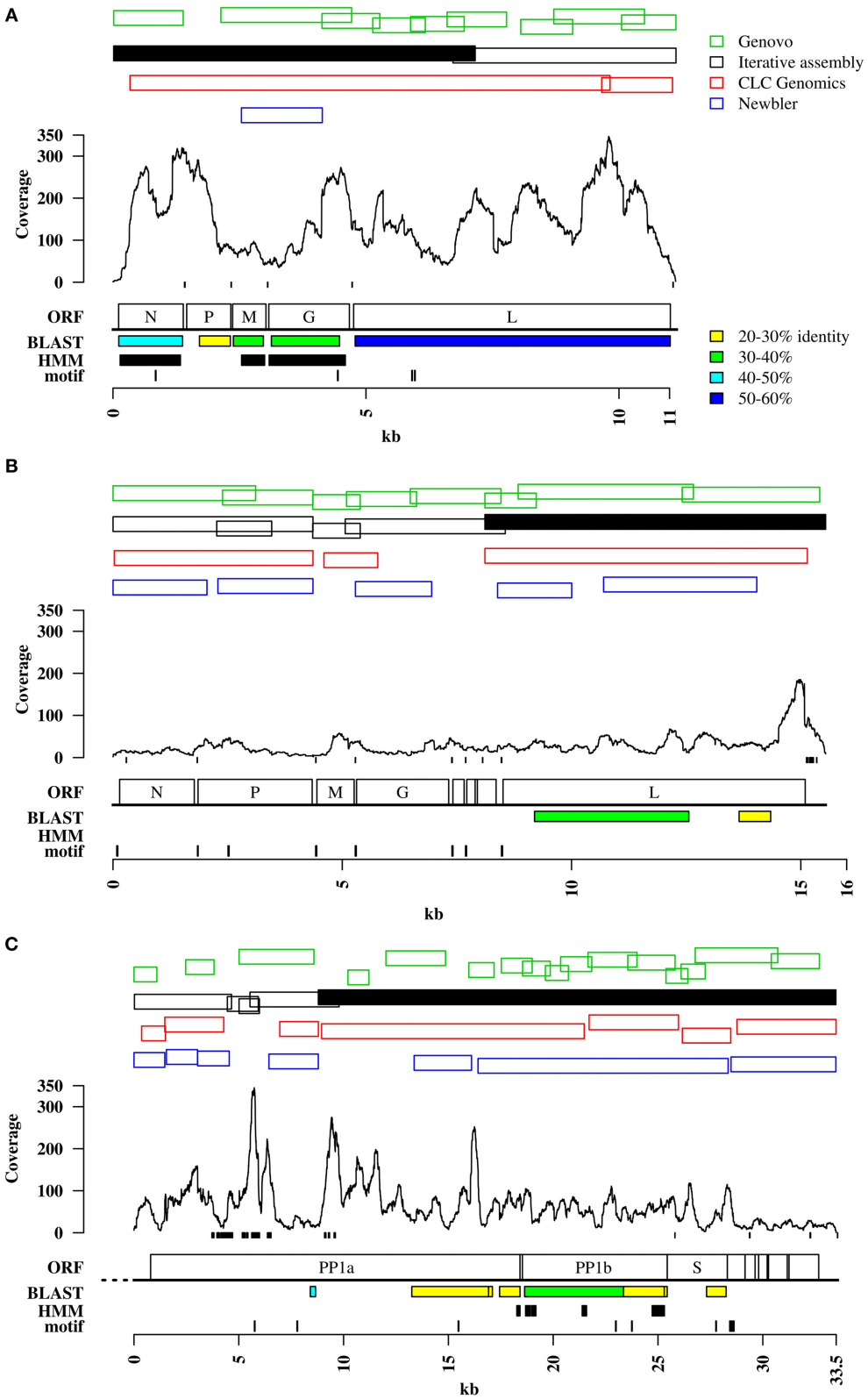


FIGURE 1 | Viral target genomes. Panels **(A–C)** contain information on read coverage and contigs matching the viral genomes of DRV **(A)**, RFFRV **(B)**, and PNV **(C)**, produced by different assembly algorithms. Shown are only contigs larger than 1 kb. Green: Contigs assembled through Genovo as described in the methods. Black outlined: Contigs assembled through iterative assembly. Black solid: Seed contig. Red: Contigs assembled through CLC Genomics workbench assembler. Blue: Contigs assembled through

(Continued)

FIGURE 1 | Continued
Newbler assembler. Small black boxes at the bottom of the read coverage line mark stretches of low sequence complexity. "ORF" indicates the genome organization as described below. "Motif" shows the location of sequence motifs. Motifs are shown in detail in **Figure S1**. "BLAST" shows regions with sequence homology as determined by BLASTX. Colored boxes show sequence identity to the best BLAST hit as indicated on top. "HMM" indicates region with remote homology identified by PFAM profiles, if any. Ruler at the bottom indicates sequence lengths in kilobases. **(A)** DRV, Dolphin rhabdovirus; N, nucleoprotein; P, phosphoprotein; M, matrix protein; G, glycoprotein; L, large protein. **(B)** RFFRV, Red fox fecal rhabdovirus; N, nucleoprotein; P, phosphoprotein; M, matrix protein; G, glycoprotein; L, large protein; no abbrevation, alpha 1,2,3 protein. **(C)** PNV, Python nidovirus; PP1a, polyprotein 1a; PP1b, polyprotein1b; S, spike glycoprotein; no abbreviations, minor membrane protein, membrane protein, nucleocapsid protein, minor membrane protein 2, putative hemagglutinin-neuraminidase protein. Striped line at 5′ end indicates putative unresolved 5′ end.

All occurrences of the detected motifs in the target genomes are indicated in **Figures 1A–C**. For both RFFRV and PNV, motif detection was able to identify contigs from genomic regions lacking BLAST or HMM-detectable homology. This information was sufficient to obtain the complete coding region of the RFFRV genome in combination with the iterative assembly approach.

COVERAGE PROFILE BINNING

In order to find additional contigs by coverage profile binning, the average coverage of every contig of the CCS and the two metagenomes was calculated by dividing the number of reads by the length of the contig. In the CCS dataset, an average coverage of 1.20 reads per base was achieved for the DRV seed contig. Accordingly, another contig that had a coverage of more than 1.1 read/base was obtained from the DRV genome (**Figure 2A**). All other contigs in the CCS dataset showed a lower coverage profile. In the RFF metagenome however, contigs with a similar coverage frequency as the seed contig of RFFRV (coverage of 0.42) were identified as putative plant genes, with a closest homolog to a hypothetical protein of *Medicago truncatula* (BLASTX *e*-value 6e-60, coverage of 0.45) or as part of the *Vulpes vulpes* mitochondrium (*e*-value 0.0, coverage of 0.58) (**Figure 2B**) and as two parts of a novel picobirnavirus, RFF picobirnavirus, isolate 40-2 (Bodewes et al., 2014b) (*e*-value 0.0, coverage of 0.299 and 0.99). In the TPLT metagenome, the seed contig of PNV was covered by 0.5 reads per base. Four of five contigs with a coverage profile of more than 0.2 were matching the PNV genome (**Figure 2C**), with the exception of one contig identified as hypothetical protein of *Clostridium thermocellum* (BLASTX *e*-value 4e-29). In conclusion, coverage profile binning identified additional contigs in two of the datasets.

K-mer FREQUENCIES

Another possible method to detect contigs that lack homology to known viruses but are part of the viral target genome is to determine k-mer frequencies. The frequency of every oligomer at length k was determined for $k = 3, 4, 5, 6, 7$ and 8 and ranked according to their absolute difference with the sum of the k-mer frequencies of the seed contig (**Figure 3**). Low ranking contigs have a similar k-mer frequency profile as the seed contig, whereas high ranking contigs differ in their k-mer frequency profile. For the CCS dataset, one 7.5 kb contig had a closely matching, high ranking k-mer frequency profile and was indeed originating from DRV (**Figure 3A**). Similarly, the two largest contigs (>3.5 kb) of RFFRV had the highest rank when compared to the frequency profile of the seed contig (**Figure 3B**). Two smaller contigs ranked at 14 and 24, suggesting that k-mer frequency profile clustering works better with long sequences. However, for PNV, the highest ranking, largest contig was obtained from the python host genome and contained among others a sequence for cytochrome C oxidase subunit (BLASTX *e*-value 0). Two large contigs matching the PNV genome ranked at 22 and 47, and two small contigs that were obtained from the PNV genome had an even higher rank (**Figure 3C**). While k-mer frequency ranking identified an additional part of DRV, and two large RFFRV contigs, all high-ranking contigs (rank 10 or less) of dataset 3 were unrelated to PNV.

DISCUSSION

We tested and compared different strategies to assist in identifying viral contigs and increasing the level of viral genome completeness from metagenomes. The retrospective nature of this study allowed us to compare the success of the strategies in retrieving three novel viral genomes from 454 metagenomic data. While different metagenome assembly strategies, especially for very large datasets of short read data, apply k-mer clustering or digital normalization and partitioning prior to assembly (Howe et al., 2014; Reddy et al., 2014), we here concentrated on strategies to link contigs after assembly *in silico*. Theoretically, these strategies can also be applied to contigs from metagenomes produced by other sequencing methods. A growing number of metagenome studies apply other next-generation sequencing techniques (e.g., Illumina), but 454 sequencing is still widely applied in viral metagenome studies, sometimes in combination with Illumina or PacBio sequences, because of the large read length (up to 800 bases) (De Vries et al., 2012; Grard et al., 2012; Philippe et al., 2013). In all three datasets presented here, the viral load of the samples was exceptionally high. Despite the high number of reads obtained from the target virus, direct assembly of the full genomes was not possible.

Metagenome assembly is a challenging task, because the number, nature, and abundance of the genomes present in the metagenome is unknown. Whole-genome assemblers are not suited for this task (Laserson et al., 2011; Peng et al., 2011; Lai et al., 2012; Namiki et al., 2012; Scholz et al., 2012). They assume even coverage and recognize high-coverage regions as repeats rather than a highly abundant species, or as an unevenly covered region introduced by amplification bias. Virus discovery relies heavily on low input material methods; therefore an amplification strategy is often necessary. Common random amplification methods, such as MDA or sequence-independent SISPA (Dean et al., 2002; Hutchison et al., 2005; Spits et al., 2006) are known to produce strong amplification biases leading to highly uneven coverage depths (Karlsson et al., 2013; Rosseel et al., 2013).

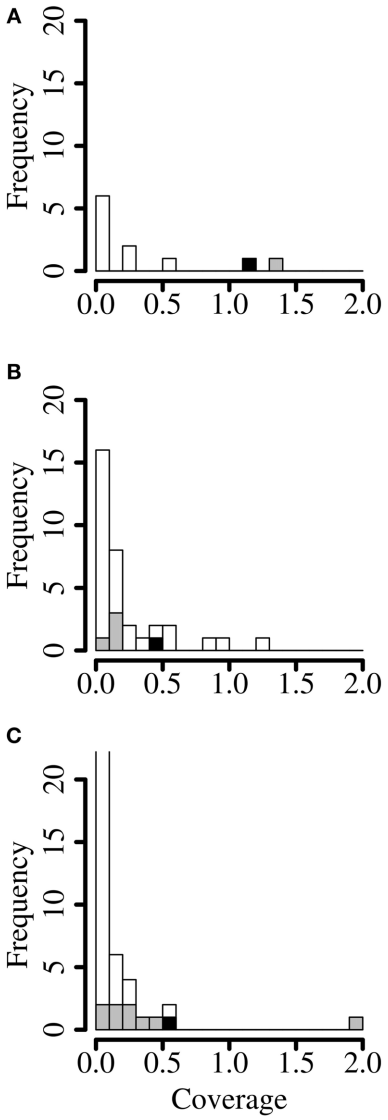


FIGURE 2 | Coverage profile binning. Histograms of coverage (reads per base) of each contig of **(A)** of cell culture supernatant containing Dolphin rhabdovirus, **(B)** red fox feces containing red fox fecal rhabdovirus and **(C)** python lung tissue containing python nidovirus. Gray: contigs mapping to the finished viral genome. Black: seed contig. The first bar in the last panel is truncated for visibility (47%). Shown are only contigs larger than 1 kb.

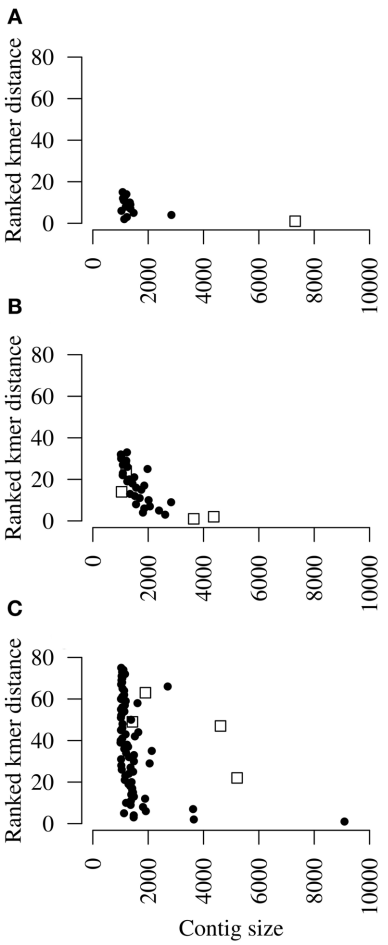


FIGURE 3 | K-mer profiling. Dot plots showing ranked k-mer distance of each contig when compared to the k-mer profile of the seed contig of **(A)** Dolphin rhabdovirus (DRV), **(B)** red fox fecal rhabdovirus (RFFRV), and **(C)** python nidovirus (PNV) in relation to contig lengths. Open boxes indicate contigs that were retrospectively identified as originating from the target genomes. Shown are only contigs larger than 1 kb.

Nevertheless the need for amplification makes these two methods still the most widely used in virus discovery (Allander et al., 2005; Djikeng et al., 2008; Hall et al., 2014). Introduction of amplification bias leads to stretches in the viral genome that are better covered than others. This can not only mislead assembly, it could also hamper detection of additional contigs by coverage profile binning. Accordingly, coverage binning was a successful strategy to link additional contigs for DRV and PNV, but not for RFFRV. Nevertheless, coverage profile binning was successfully applied to assemble viral genomes across a number of human gut metagenomes without the need of a reference (Nielsen et al., 2014) or to verify a cross-assembly of a novel bacteriophage in similar samples (Dutilh et al., 2014). Due to availability of a large amount of (fecal) sample for metagenome studies, amplification can often be avoided, which makes the application of coverage profile binning more straight-forward.

An additional issue in the case of viral metagenomes is the presence of distinct quasispecies sequences which can hamper direct assembly, especially at low sequencing depths. Using stringent assembly parameters that are necessary to avoid chimeras can lead to highly similar singletons or small contigs that are too diverse for assembly into a population sequence. This problem can be overcome by reference-guided assembly by a quasispecies assembler (Prosperi et al., 2013). However, this is currently not possible for divergent viruses which lack a reference genome. While many metagenome assemblers to date were designed to handle short-read data, only very few assemblers are dedicated to assembly of longer (i.e., 454) reads without any further

information such as paired-end or mate-pair information. For this study, we used different assemblers, including an overlap-layout consensus algorithm (Newbler), an assembler that uses a generative probabilistic model (Genovo; Laserson et al., 2011), a de Bruijn graph algorithm (CLC Genomics Workbench), and an assembly strategy applying the combination of an OLC and a greedy algorithm (iterative assembly; Schurch et al., 2014). None of these strategies lead to a full reconstruction of the genomes of the novel viruses, but produced fragmented contigs. Overlaps between the contigs were often not recognized because of misassembled contig ends (not indicated in **Figures 1A–C**). These misassembled ends could represent chimeric contigs, i.e., assembled from reads from different species, but also chimeric reads due to chimera formation during PCR. Chimerism can not only prohibit successful assembly but can also lead to misclassification of the taxonomic content of the metagenome sample (Mavromatis et al., 2007; Pignatelli and Moya, 2011; Mende et al., 2012). Taxonomic "misclassification" of reads in the analysis described here, however, was rather due to the large number of taxonomic units without a homolog in the sequence databases. These reads were then classified as "unknowns." Another challenge for recovery of viral genomes from metagenomes poses the segmented genomes of some viruses, with up to 12 segments for some viruses in the family *Reoviridae*, for example the Colorado tick fever virus (Attoui et al., 2000). Those segments can only be separately assembled, if possible, and need to be linked afterwards. The strategies described in this study can aid in identification of missing segments or contigs.

The strategy with the highest specificity was *de novo* motif discovery in the seed contig, and subsequent motif search in all contigs of the assembled metagenome. The (A/U)CU7 motif detected between open reading frames of RFFRV could serve as a transcription termination/polyadenylation sequence similar to other rhabdoviruses (Whelan et al., 2004).

Adjacent to this termination signal was a stretch of conserved nucleotides which might function as a transcription initiation signal. For the other detected motifs in DRV and PNV no obvious functions can be envisaged. However, their power to detect additional contigs matching the target genomes was only limited by the number of occurrences of the motif in the genome.

K-mer profile ranking detected large viral contigs with a similar profile as the seed contig in the CCS dataset and the RFF metagenome. In both cases, further manual curation or a laboratory follow-up would have been necessary to confirm the predictions made by this technique.

Assembly, in combination with motif discovery enabled retrieval of the complete RFFRV genome, with good results in k-mer frequency clustering. Two additional contigs of the PNV genome were identified by motif search, but linking of the remaining PNV contigs was only possible with frequency methods. Surprisingly, all methods applied here showed good results in retrieval of the full genome of DRV from the CCS dataset. This is most likely due to the high viral load which allowed assembly of the whole genome into two very long contigs in the first place. Therefore, we feel that the development of more efficient and dedicated metagenome assemblers, taking into account the specific characteristics of viral genomes, will lead to improved retrieval of viral pathogen genomes from metagenome sequences.

In conclusion, iterative exhaustive assembly, although highly stringent and thus excluding a large amount of data, is actually performing rather well compared to other assembly algorithms in that it covered the target genomes more completely than any other set of contigs obtained with other assembly algorithms. Nevertheless, the number of identified target virus reads and the level of viral genome completeness can be increased by combining data generated with different assembly algorithms. In addition, various methods can be applied to obtain additional genome fragments although a success rate cannot be predicted beforehand based on our analyses and probably depend largely on the dataset under study. These results indicate that a combination of these methods can be of great value to rapidly obtain additional genome information of a previously unknown virus.

AUTHOR CONTRIBUTIONS

Rogier Bodewes and Anita C. Schürch conceived the study. Anita C. Schürch designed the experiments. Anita C. Schürch, Rogier Bodewes carried out the research. Saskia L. Smits contributed to the design of experiments. Anita C. Schürch prepared the first draft of the manuscript. Rogier Bodewes, Saskia L. Smits, Aritz Ruiz-Gonzalez, Wolfgang Baumgärtner, contributed materials. Saskia L. Smits, Marion P. Koopmans, Albert D. M. E. Osterhaus participated in the discussion and writing of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

GRANT INFORMATION

This work was partially funded by the Virgo Consortium, funded by the Dutch government project number FES0908, by Netherlands Genomics Initiative (NGI) project number 050-060-452 and ZonMW TOP project 91213058. A. Ruiz-Gonzalez holds a Post doc fellowship awarded by the Department of Education, Universities and Research of the Basque Government (Ref. DKR-2012-64) and was partially supported by the Research group on "Systematics, Biogeography and Population Dynamics" (Basque Government; Ref. IT317-10; GIC10/76).

ACKNOWLEDGMENTS

The authors wish to thank Jurre Y. Siegers for characterization of the DRV genome.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2014.00714/abstract

Figure S1 | (A) Nucleotide sequence motif discovered in the Dolphin rhabdovirus (DRV). **(B)** Nucleotide sequence motif discovered in the red fox fecal rhabdovirus (RFFRV). **(C)** Nucleotide sequence motif discovered in the python nidovirus (PNV).

REFERENCES

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

Allander, T., Tammi, M. T., Eriksson, M., Bjerkner, A., Tiveljung-Lindell, A., and Andersson, B. (2005). Cloning of a human parvovirus by molecular screening of respiratory tract samples. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12891–12896. doi: 10.1073/pnas.0504666102

Attoui, H., Billoir, F., Cantaloube, J. F., Biagini, P., De Micco, P., and De Lamballerie, X. (2000). Strategies for the sequence determination of viral dsRNA genomes. *J. Virol. Methods* 89, 147–158. doi: 10.1016/S0166-0934(00)00212-3

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Bloch, K. C., and Glaser, C. (2007). Diagnostic approaches for patients with suspected encephalitis. *Curr. Infect. Dis. Rep.* 9, 315–322. doi: 10.1007/s11908-007-0049-5

Bodewes, R., Lempp, C., Schurch, A. C., Habierski, A., Hahn, K., Lamers, M., et al. (2014a). Novel divergent nidovirus in a python with pneumonia. *J. Gen. Virol.* 95, 2480–2485. doi: 10.1099/vir.0.068700-0

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C. M., Van Den Brand, J. M., Osterhaus, A. D., and Smits, S. L. (2014b). Viral metagenomic analysis of feces of wild small carnivores. *Virol. J.* 11:89. doi: 10.1186/1743-422X-11-89

Bodewes, R., Ruiz-Gonzalez, A., Schürch, A. C., Osterhaus, A. D., and Smits, S. L. (2014c). Novel rhabdovirus in feces of a red fox, Spain. *Emerg. Infect. Dis*. 20, 2172–2174. doi: 10.3201/eid2012.140236

Bodewes, R., Van De Bildt, M. W., Schapendonk, C. M., Van Leeuwen, M., Van Boheemen, S., De Jong, A. A., et al. (2013). Identification and characterization of a novel adenovirus in the cloacal bursa of gulls. *Virology* 440, 84–88. doi: 10.1016/j.virol.2013.02.011

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

Chew, Y. V., and Holmes, A. J. (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *J. Microbiol. Methods* 78, 136–143. doi: 10.1016/j.mimet.2009.05.003

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., et al. (2002). Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5261–5266. doi: 10.1073/pnas.082089499

Delwart, E. L. (2007). Viral metagenomics. *Rev. Med. Virol.* 17, 115–131. doi: 10.1002/rmv.532

Denno, D. M., Shaikh, N., Stapp, J. R., Qin, X., Hutter, C. M., Hoffman, V., et al. (2012). Diarrhea etiology in a pediatric emergency department: a case control study. *Clin. Infect. Dis.* 55, 897–904. doi: 10.1093/cid/cis553

De Vries, M., Oude Munnink, B. B., Deijs, M., Canuti, M., Koekkoek, S. M., Molenkamp, R., et al. (2012). Performance of VIDISCA-454 in feces-suspensions and serum. *Viruses* 4, 1328–1334. doi: 10.3390/v4081328

Djikeng, A., Halpin, R., Kuzmickas, R., Depasse, J., Feldblyum, J., Sengamalay, N., et al. (2008). Viral genome sequencing by random priming methods. *BMC Genomics* 9:5. doi: 10.1186/1471-2164-9-5

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5:4498. doi: 10.1038/ncomms5498

Garcia-Garcera, M., Garcia-Etxebarria, K., Coscolla, M., Latorre, A., and Calafell, F. (2013). A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PLoS ONE* 8:e74914. doi: 10.1371/journal.pone.0074914

Grard, G., Fair, J. N., Lee, D., Slikas, E., Steffen, I., Muyembe, J. J., et al. (2012). A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog.* 8:e1002924. doi: 10.1371/journal.ppat.1002924

Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *J. Virol. Methods* 195, 194–204. doi: 10.1016/j.jviromet.2013.08.035

Handley, K. M., Bartels, D., O'Loughlin, E. J., Williams, K. H., Trimble, W. L., Skinner, K., et al. (2014). The complete genome sequence for putative H - and S-oxidizer Candidatus Sulfuricurvum sp., assembled *de novo* from an aquifer-derived metagenome. *Environ. Microbiol.* 16, 3443–3462. doi: 10.1111/1462-2920.12453

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proc. Natl. Acad. Sci. U.S.A.* 111, 4904–4909. doi: 10.1073/pnas.1402564111

Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868–877. doi: 10.1101/gr.9.9.868

Hutchison, C. A. 3rd., Smith, H. O., Pfannkoch, C., and Venter, J. C. (2005). Cell-free cloning using phi29 DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.* 102, 17332–17336. doi: 10.1073/pnas.0508809102

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., and Armbrust, E. V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335, 587–590. doi: 10.1126/science.1212665

Karlsson, O. E., Hansen, T., Knutsson, R., Lofstrom, C., Granberg, F., and Berg, M. (2013). Metagenomic detection methods in biopreparedness outbreak scenarios. *Biosecur. Bioterror.* 11(Suppl 1), S146–S157. doi: 10.1089/bsp.2012.0077

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., et al. (2012). Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* 22, 292–298. doi: 10.1101/gr.126573.111

Lai, B., Ding, R., Li, Y., Duan, L., and Zhu, H. (2012). A *de novo* metagenomic assembly program for shotgun DNA reads. *Bioinformatics* 28, 1455–1462. doi: 10.1093/bioinformatics/bts162

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: *de novo* assembly for metagenomes. *J. Comput. Biol.* 18, 429–443. doi: 10.1089/cmb.2010.0244

Lipkin, W. I. (2010). Microbe hunting. *Microbiol. Mol. Biol. Rev.* 74, 363–377. doi: 10.1128/MMBR.00007-10

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* 4, 495–500. doi: 10.1038/nmeth1043

Mende, D. R., Waller, A. S., Sunagawa, S., Jarvelin, A. I., Chan, M. M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* 7:e31386. doi: 10.1371/journal.pone.0031386

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40, e155. doi: 10.1093/nar/gks678

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* 32, 822–828. doi: 10.1038/nbt.2939

Osterhaus, A. D., Broeders, H. W., Teppema, J. S., Kuiken, T., House, J. A., Vos, H. W., et al. (1993). Isolation of a virus with rhabdovirus morphology from a white-beaked dolphin (*Lagenorhynchus albirostris*). *Arch. Virol.* 133, 189–193. doi: 10.1007/BF01309754

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (in press). *Biostrings: String Objects Representing Biological Sequences, and Matching Algorithms.* R package version 2.32.30. Available online at: http://www.bioconductor.org/packages/release/bioc/html/Biostrings.html

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., et al. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* 341, 281–286. doi: 10.1126/science.1239181

Pignatelli, M., and Moya, A. (2011). Evaluating the fidelity of *de novo* short read metagenomic assembly using simulated data. *PLoS ONE* 6:e19984. doi: 10.1371/journal.pone.0019984

Prachayangprecha, S., Schapendonk, C. M., Koopmans, M. P., Osterhaus, A. D., Schurch, A. C., Pas, S. D., et al. (2014). Exploring the potential of next-generation sequencing in diagnosis of respiratory viruses. *J. Clin. Microbiol.* 52, 3722–3730. doi: 10.1128/JCM.01641-14

Prosperi, M. C., Yin, L., Nolan, D. J., Lowe, A. D., Goodenow, M. M., and Salemi, M. (2013). Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci. Rep.* 3:2837. doi: 10.1038/srep02837

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucl. Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Reddy, R. M., Mohammed, M. H., and Mande, S. S. (2014). MetaCAA: a clustering-aided methodology for efficient assembly of metagenomic datasets. *Genomics* 103, 161–168. doi: 10.1016/j.ygeno.2014.02.007

Rosseel, T., Van Borm, S., Vandenbussche, F., Hoffmann, B., Van Den Berg, T., Beer, M., et al. (2013). The origin of biased sequence depth in sequence-independent nucleic acid amplification and optimization for efficient massive parallel sequencing. *PLoS ONE* 8:e76144. doi: 10.1371/journal.pone.0076144

Schmieder, R., and Edwards, R. (2012). Insights into antibiotic resistance through metagenomic approaches. *Future Microbiol.* 7, 73–89. doi: 10.2217/fmb.11.135

Scholz, M. B., Lo, C.-C., and Chain, P. S. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* 23, 9–15. doi: 10.1016/j.copbio.2011.11.013

Schurch, A. C., Schipper, D., Bijl, M. A., Dau, J., Beckmen, K. B., Schapendonk, C. M., et al. (2014). Metagenomic survey for viruses in western arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS ONE* 9:e105227. doi: 10.1371/journal.pone.0105227

Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* 5:209. doi: 10.3389/fpls.2014.00209

Siegers, J. Y., Van De Bildt, M. W., Van Elk, C. E., Schurch, A. C., Tordo, N., Kuiken, T., et al. (2014). Genetic relatedness of dolphin rhabdovirus with fish rhabdoviruses. *Emerging Infect. Dis.* 20, 1081–1082. doi: 10.3201/eid2006.131880

Smits, S. L., and Osterhaus, A. D. (2013). Virus discovery: one step beyond. *Curr. Opin. Virol.* doi: 10.1016/j.coviro.2013.03.007. [Epub ahead of print].

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., et al. (2006). Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* 1, 1965–1970. doi: 10.1038/nprot.2006.326

Team, R. C. (2012). *R: A Language and Environment for Statistical Computing*. Available online at: http://web.mit.edu/r_v3.0.1/fullrefman.pdf

Van Den Brand, J. M., Van Leeuwen, M., Schapendonk, C. M., Simon, J. H., Haagmans, B. L., Osterhaus, A. D., et al. (2012). Metagenomic analysis of the viral flora of pine marten and European badger feces. *J. Virol.* 86, 2360–2365. doi: 10.1128/JVI.06373-11

Van Leeuwen, M., Williams, M. M., Koraka, P., Simon, J. H., Smits, S. L., and Osterhaus, A. D. (2010). Human picobirnaviruses identified by molecular screening of diarrhea samples. *J. Clin. Microbiol.* 48, 1787–1794. doi: 10.1128/JCM.02452-09

Whelan, S., Barr, J., and Wertz, G. (2004). "Transcription and replication of non-segmented negative-strand RNA viruses," in *Biology of Negative Strand RNA Viruses: The Power of Reverse Genetics*, ed Y. Kawaoka (Berlin; Heidelberg: Springer), 61–119.

Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443, 950–955. doi: 10.1038/nature05192

Wylie, K. M., Truty, R. M., Sharpton, T. J., Mihindukulasuriya, K. A., Zhou, Y., Gao, H., et al. (2012). Novel bacterial taxa in the human microbiome. *PLoS ONE* 7:e35294. doi: 10.1371/journal.pone.0035294

Conflict of Interest Statement: Drs. Albert D. M. E. Osterhaus and Saskia L. Smits are partly employed by Viroclinics Biosciences B.V., Rotterdam, Netherlands. The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 17 September 2014; paper pending published: 20 October 2014; accepted: 30 November 2014; published online: 18 December 2014.

Citation: Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgärtner W, Koopmans MP, Osterhaus ADME and Schürch AC (2014) Assembly of viral genomes from metagenomes. Front. Microbiol. ***5****:714. doi: 10.3389/fmicb.2014.00714*

This article was submitted to Virology, a section of the journal Frontiers in Microbiology.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 10 April 2015
doi: 10.3389/fmicb.2015.00265

Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions

Cheryl-Emiliane T. Chow[1], *Danielle M. Winget*[1†], *Richard A. White III*[2†], *Steven J. Hallam*[2,3,4] *and Curtis A. Suttle*[1,2,3,5*]

[1] Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada, [2] Department of Microbiology and Immunology, University of British Columbia, Vancouver, BC, Canada, [3] Integrated Microbial Biodiversity Program, Canadian Institute for Advanced Research, Toronto, ON, Canada, [4] Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC, Canada, [5] Department of Botany, University of British Columbia, Vancouver, BC, Canada

Viral diversity and virus-host interactions in oxygen-starved regions of the ocean, also known as oxygen minimum zones (OMZs), remain relatively unexplored. Microbial community metabolism in OMZs alters nutrient and energy flow through marine food webs, resulting in biological nitrogen loss and greenhouse gas production. Thus, viruses infecting OMZ microbes have the potential to modulate community metabolism with resulting feedback on ecosystem function. Here, we describe viral communities inhabiting oxic surface (10 m) and oxygen-starved basin (200 m) waters of Saanich Inlet, a seasonally anoxic fjord on the coast of Vancouver Island, British Columbia using viral metagenomics and complete viral fosmid sequencing on samples collected between April 2007 and April 2010. Of 6459 open reading frames (ORFs) predicted across all 34 viral fosmids, 77.6% ($n = 5010$) had no homology to reference viral genomes. These fosmids recruited a higher proportion of viral metagenomic sequences from Saanich Inlet than from nearby northeastern subarctic Pacific Ocean (Line P) waters, indicating differences in the viral communities between coastal and open ocean locations. While functional annotations of fosmid ORFs were limited, recruitment to NCBI's non-redundant "nr" database and publicly available single-cell genomes identified putative viruses infecting marine thaumarchaeal and SUP05 proteobacteria to provide potential host linkages with relevance to coupled biogeochemical cycling processes in OMZ waters. Taken together, these results highlight the power of coupled analyses of multiple sequence data types, such as viral metagenomic and fosmid sequence data with prokaryotic single cell genomes, to chart viral diversity, elucidate genomic and ecological contexts for previously unclassifiable viral sequences, and identify novel host interactions in natural and engineered ecosystems.

Keywords: microbial ecology, marine virus, metagenomics, fosmids, virome, reference genome, single cell genomics

OPEN ACCESS

Edited by:
Richard J. Hall, Institute of Environmental Science and Research, New Zealand

Reviewed by:
Jianwei Wang, Chinese Academy of Medical Sciences & Peking Union Medical College, China
F. Murilo Zerbini, Universidade Federal de Viçosa, Brazil

***Correspondence:**
Curtis A. Suttle, Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Rm. 2020, 2207 Main Mall, Vancouver, BC V6T 1Z4, Canada
suttle@science.ubc.ca

[†]***Present Address:***
Danielle M. Winget, Department of Biology, Seattle Pacific University, Seattle, USA; Richard A. White III, Pacific Northwest National Laboratory, Richland, USA

Specialty section:
This article was submitted to Virology, a section of the journal Frontiers in Microbiology

Received: *16 January 2015*
Accepted: *17 March 2015*
Published: *10 April 2015*

Citation:
Chow C-ET, Winget DM, White RA III, Hallam SJ and Suttle CA (2015) Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. Front. Microbiol. 6:265. doi: 10.3389/fmicb.2015.00265

Introduction

The long evolutionary history of viruses with cellular life is evident from the diseases they cause, such as influenza and AIDS, and also from the viral genes found in the genomes of cells. These relationships have their origins in viruses that infect bacteria, archaea and protists, all of which play a critical role in global nutrient and energy cycling and in maintaining functional ecosystems. Viruses affect the abundance and diversity of phytoplankton (e.g., Larsen et al., 2004), bacteria (e.g., Winter et al., 2004), and archaea (e.g., Andersson and Banfield, 2008), and consequently influence global biogeochemical cycles (Fuhrman, 1999; Wilhelm and Suttle, 1999; Suttle, 2007; Winget et al., 2011), and genome evolution (Shackelton and Holmes, 2004; Sharon et al., 2007). Despite our emerging understanding that viruses play important roles in the earth system, our knowledge of the distribution of viral genotypes, their dispersal among environments, their ecological niches, and the functions of most viral genes remain largely unknown (Brussaard et al., 2008).

Although advances in nucleic-acid technologies have greatly increased the rate and depth to which the genetic diversity and ecology of viral communities can be interrogated, the inferences drawn from sequence-based investigations are hampered by methodological biases and non-representative databases of viral sequences. These limitations are exacerbated by the ultra-low quantity of nucleic acids in viral particles, the enormous genetic diversity of viruses in nature, and the lack of relevant model systems across a breadth of viral taxonomic groups. For example, most of the sequenced and available dsDNA viral genomes are from tailed phages in the order *Caudovirales*, although these and related genotypes are not dominant in marine systems (e.g., Breitbart et al., 2002; Angly et al., 2006). Consequently, only a small proportion of viral metagenomic reads can be aligned with sequenced viral genomes and placed in a genomic context. Moreover, the majority of predicted viral open reading frames (ORFs) have no functional annotation (Angly et al., 2006; Williamson et al., 2012; Hurwitz and Sullivan, 2013; Hurwitz et al., 2015), leaving viral ecologists to wonder what most of this genetic material represents. In the absence of an abundance of and wider diversity of viral isolates and host systems, viral genomic information must be gleaned using alternative methods [e.g., large insert fosmid libraries (Garcia-Heredia et al., 2012; Mizuno et al., 2013b), sequencing of viral DNA extracted from pulsed-field gel electrophoresis bands (Ray et al., 2012), or single-virus genomics (Allen et al., 2011)]. Additionally, by targeting and sorting viral or host populations with flow cytometry, genomic data can be obtained for specific virus-host interactions (Deng et al., 2014; Martinez-Martinez et al., 2014). Mining cellular metagenomic and single-cell genome datasets has also unearthed new virus genomes and identified potential virus-host relationships (Anantharaman et al., 2014; Roux et al., 2014b) from previously uncultured hosts. These inferred virus-host interactions not only reveal a past virus encounter and subsequent infection of a host organism but also indicate the potential for genetic exchange during the infection cycle that can drive consequent effects on the metabolic status and rates of the infected host. When viral genomic data can be linked to a specific host organism, it becomes possible to study virus-host interactions within natural or engineered ecosystems and place "viral dark matter" into an ecological context.

In this study, viral fosmid and metagenomic sequences combined with bacterial single-cell genomes (SAGs) were interrogated with the goals of placing viral metagenomic sequence data into a genomic context and revealing host-virus interactions. Large-insert fosmid sequences (~35 kb) served as proxies for partial or nearly-complete dsDNA viral genomes; for the most abundant viruses, typical genome sizes range from about 29 to 69 kb in seawater (Steward et al., 2000). Saanich Inlet, British Columbia was used as a model site as it is a fjord that undergoes seasonal cycles of stratification and renewal that dynamically alter the oxygenation status of the water column (Anderson and Devol, 1973). During peak stratification, a redoxcline develops with anoxic and sulfidic conditions prevailing in the deep basin waters (200 m). Expansion of areas of low oxygen concentration are becoming of increasing concern worldwide (Wright et al., 2012), yet viruses and their roles in these low oxygen marine environments remain poorly studied. Comparison of metagenomic and fosmid sequences show that viruses in Saanich Inlet were distinct from those in other environments and identified putative viruses infecting marine thaumarchaea and members of the bacterial SUP05 clade.

Materials and Methods

Sample Collection

Sample collection was carried out on board the *MSV John Strickland* in Saanich Inlet, British Columbia at station S3 (48° 35′ 30.0012″N, 123° 30′ 21.9996″W). A sill at the fjord mouth prevents mixing and oxygenation except for deep water renewal events in early September and after unusually strong storms (Anderson and Devol, 1973), which leads to hypoxic condition below the mixed layer for most of the year. Approximately 20 L of seawater was collected monthly by wire-mounted Niskin bottles at 10 m and 200 m depth intervals from April 2007 to April 2010. Seawater was filtered through a 0.22-μm pore-size Sterivex filter (Millipore) to remove the cellular fraction. Viruses were concentrated by tangential flow ultrafiltration through a 30-kDa molecular-weight cutoff cartridge (Prep-Scale 2.5, Millipore) to a final volume between 250 and 500 mL (Suttle et al., 1991), and stored at 4°C until further processing.

Ten-mL subsamples of viral concentrates (VCs) from either 10 m or 200 m were combined into composite mixes to create fosmid libraries for each of summer, fall, and winter and a mix of 32 VCs spanning 3 years was used for metagenomic sequencing (Table S1). VC mixes were filtered again through 47 mm diameter, 0.22-μm pore-size filters after mixing (Type PVDF: polyvinylidene fluoride, Millipore) and then further concentrated to between 1 and 2 mL using a 30 kDa molecular-weight cutoff Centricon filter by spinning at 3000 rpm (~825 × g) for 8–10 min at 10°C in a benchtop centrifuge.

DNA Extraction and Sequencing Library Preparation

Viral Fosmid Libraries

Within 48 h of Centricon concentration, viral DNA was extracted from each VC mix as follows. First, free DNA and RNA were removed by incubation with 1 μl each of DNAse I (1 U/μl) and RNAse A (20 mg/ml) in a final concentration of 1x DNAse reaction buffer (Invitrogen) for 15 min at room temperature. Enzymes were inactivated by addition of 1 μL of 25 mM ETDA and incubation at 65°C for 10 min. DNA was then extracted in multiple 50 μL aliquots using the Gentra Puregene Blood kit (Qiagen) per the manufacturer's recommendations. Samples were subjected to Proteinase K treatment and repeated protein precipitation steps as advised by the manufacturer. Final DNA extracts were rehydrated in 10 μL of sterile DNAse- and RNAse-free water (Gibco) at 4°C overnight to elute DNA pellets. The 10 μL DNA extracts from each of the 50 μl VC aliquots were pooled, and the isopropanol and ethanol precipitation steps were repeated to further concentrate DNA. The final DNA sample was again eluted in sterile water and stored at −20°C.

Fosmid cloning was performed using the CopyControl Fosmid Library Production Kit (Epicentre) according to manufacturer's protocols with the following modifications. End-repaired DNA was immediately ligated without further size selection to avoid loss of material. Ligation of DNA into the CopyControl vector occurred overnight at 16°C. Twelve to twenty colonies were picked for each seasonal VC mix and grown overnight in selective media (LB + 12.5 μg mL^{-1} chloramphenicol) with addition of CopyControl Fosmid Autoinduction Solution at 1X final concentration to induce high copy number production of the fosmid vector. Fosmid DNA was purified from the high copy number induced overnight cultures using the FosmidMax DNA Purification Kit (Epicenter) and stored at −20°C. Glycerol stocks of overnight cultures were stored at −80°C.

To assess fosmid insert size and genetic differences, 2 μL of purified fosmid DNA was digested with Apa I at 25°C for 2 h followed by inactivation at 65°C for 20 min and visualized by pulsed-field gel electrophoresis (1% low melting point agarose gel, 0.5x TBE, 14°C, voltage gradient 6.0 V cm^{-1}, total run time 22 h, initial switch time 1 s, final switch time 15 s, linear ramping factor). Each of six samples for sequencing (3 seasons × 2 depths) was composed of 2 μg of DNA from each of the 12 fosmids selected per sample based on differences in restriction digest patterns. Samples were sequenced using 454 Titanium chemistry (Cambridge, MA, USA).

Viral Metagenomes

Viral DNA was extracted following the concentration of the 32 VCs by 30 kDa Centricon ultrafiltration into a single sample per depth (10 m and 200 m) and treatment to remove free DNA and RNA as stated above. Each sample was then divided in half for parallel DNA extraction. Two DNA extraction kits were used to minimize any potential bias in genomic extraction by either kit. Half the sample was extracted using Gentra Puregene Blood kit (Qiagen) as per the manufacturer's recommendations with the same modifications as listed for fosmid library DNA extraction. The second half of the sample was extracted using the QIAamp Virus MinElute Spin Kit (Qiagen) according to the manufacturer's protocol. Extracted DNA was frozen at −20°C until further processing. DNA from both protocols was thawed and combined just prior to multiple displacement amplification with random primers, per manufacturer's instructions (GenomePlex Complete Whole Genome Amplification kit, Sigma-Aldrich Canada Co, Oakville, Ontario, Canada), to increase DNA amounts prior to library construction. Amplified DNA was pooled within each depth to reflect a composite community to minimize seasonal biases.

For library construction, DNA was sheared by ultrasonication (Covaris M220 series, Woburn, MA) to approximately 250–300 bp. Sheared fragments were end-repaired, A-tailed and ligated to custom TruSeq adapters (IDT, Coralville, Iowa) using the NxSeq DNA Sample Prep Kit 2 for Illumina (Lucigen, Middleton, WI). After ligation of custom TruSeq adapters, an added heat-kill step (65°C for 20 min) was used to stop ligation; then, small fragments and adapter dimers were removed twice using Agencourt AMPureXP SPRI magnetic beads (Beckman Coulter, Danvers, MA). Libraries were checked for size and adapter dimers using a High Sensitivity DNA chip on a Bioanalyzer 2100 (Agilent) and quantified using Qubit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Libraries were sequenced for 2 × 250 bp paired-end reads on an Illumina MiSeq v2.0 at the Génome Québec Innovation Centre at McGill University (Montréal, Québec, Canada).

Sequence Analysis

Viral Metagenomic Data

Metagenomic sequences were trimmed for low quality base pairs and any residual adapter sequences using the default settings for Trimmomatic v0.30 (Bolger et al., 2014). All phiX reads from the control library were removed by mapping reads to the reference genome using the bowtie2 plugin (Langmead and Salzberg, 2012) in Geneious v7.1 (created by Biomatters and available from http://www.geneious.com/). All unassembled, paired reads were merged with FLASH using default minimum overlap settings (Magoč and Salzberg, 2011). The final "reads" dataset included all merged paired-end reads and all forward reads greater than 200 bp from the remaining non-overlapping sequence pairs. The unpaired reverse reads tended to be of poorer quality and were omitted from further analysis to avoid overestimation of sequence diversity. Sequence reads were annotated by BLASTx comparison to the complete viral protein RefSeq database (release 66, as of 10 July 2014) using MetaVir (Roux et al., 2011, 2014b). The taxonomic assignments and estimated community compositions were determined using the "Genome relative Abundance and Average Size GAAS" software package (Angly et al., 2009) implemented within MetaVir, which normalizes the distribution results based on reference viral genome length and weighs the similarity significance across multiple BLAST hits. The GAAS-derived community compositions and BLASTx differences with a bitscore cutoff of 50 were used for cross-sample comparisons. Sequences are available under the project "Saanich Inlet" and sample reads are designated as Saanich_10m_r200 ($SI.10_m$) and Saanich_200m_r200 ($SI.200_m$).

Rarefaction curves were compared by subsampling 50,000 sequences and determining sequence clusters at three nucleotide sequence similarity cutoffs (75, 90, 95%) in MetaVir. A dendrogram was calculated from BLAST-based comparison and clustered by overall similarity between Saanich Inlet and other publicly available viral metagenomes. Only metagenomes with more than 50000 sequences available in MetaVir were included in the cluster analysis (pvclust, R package; MetaVir).

Viral Fosmid Libraries

Fosmid sequences were assembled using the GS *De novo* Assembler (Newbler v2.5, Roche 454 Life Sciences) with default parameters. Vector sequences were trimmed and host *E. coli* sequences were screened for and removed from assemblies using GS *De novo* Assembler. Thirty-four fosmids larger than 30 kb were retained for further annotation and analysis, including six fosmids (SI.Prokaryotic) identified as virus-like sequences from a prior fosmid library from Saanich Inlet (Walsh et al., 2009). ORFs were called using Glimmer and Genemark plugins for Geneious v5.6, allowing for a minimum length of 150 bp and overlapping ORFs on either strand. ORFs were translated and annotated by searching against NCBI's "nr" database, as of 20 July 2014, by BLASTp with a minimum *e*-value of 10^{-5}. Fosmid ORF annotations were verified by comparison against results from RAST (Aziz et al., 2008; Overbeek et al., 2014) and ACLAME (Leplae et al., 2009) by BLASTp searches using default settings and a minimum *e*-value of 10^{-5}. Fosmids were also annotated as a contig project using MetaVir (project: Saanich Inlet; sample: Saanich_fosmids) by querying against the viral RefSeq database (release 66, 10 July 2014). Fosmids were aligned with specific viral reference genomes using the progressiveMAUVE plugin for Geneious v7.1 to determine sequence homology across an entire genome or fosmid sequence (Darling et al., 2010). Fosmid sequences were deposited in Genbank (KR029577-KR029610) and to CAMERA under the Moore Marine Phage/Virus Metagenomes as CAM_SMPL_000964 (Oxic_3), CAM_SMPL_000965 (Anoxic_3), CAM_SMPL_000971 (An oxic_1); CAM_SMPL_000982 (Oxic_2), CAM_SMPL_000989 (Anoxic_2), and CAM_SMPL_000993 (Oxic_1).

Comparative Fosmid, Metagenomic, and Single-Cell Genome Sequence Analysis

Metagenomic reads from Saanich Inlet ($SI.10_m$, $SI.200_m$) and Line P [from the Pacific Ocean Virome (POV) dataset (Hurwitz and Sullivan, 2013)] were queried by BLASTn against a custom database containing all viral genomes from RefSeq (release 66) and viral fosmid sequences from the Mediterranean Sea (Mizuno et al., 2013b) and Saanich Inlet (this study). For this analysis, only hits with a maximum *e*-value of 10^{-5}, greater than 50 bp in alignment length, and greater than 90% nucleotide identity were considered significant to minimize potential error. Line P viral metagenomes, specifically, were queried to determine presence and relative abundance of SI fosmids in waters with similar environmental characteristics (Wright et al., 2012). Additional viral metagenomes from CAMERA (Table S2) were queried against Saanich Inlet fosmids by BLASTn using an *e*-value cutoff of 10^{-5}. Long reference sequences provide more template and opportunity for read recruitment than short ones. As multiple reference genomes of varying length were included in RefSeq, and many are significantly larger than the average fosmid length from our study and in the Mediterranean Sea project (the databases under comparison), read recruitment to the reference databases were normalized according to reference genome size (per kbp) and metagenome size (per Gbp).

Regions of genetic similarity between the fosmid Oxic1_7 and the reverse complement of the putative archaeal provirus Pro_Nvie1 were determined by aligning both sequences with tBLASTx. Regions with an *e*-value less than 10^{-5} were plotted with genoPlotR in R (Guy et al., 2010) and included ORF annotations when available.

$SI.10_m$ and $SI.200_m$ sequences were recruited individually against viral fosmids (this study) and selected single-cell genomes (SAG) by bowtie2 using local recruitment and "high-sensitivity" in Geneious v7.1 resulting in recruitment of only reads at greater than 90% nucleotide identity. In brief, the SAG datasets used here originated from whole genome amplification and sequencing of single cells. These datasets were selected for their relevance to our study location [i.e., SAGs from the same location (Roux et al., 2014a)] and potential for discovery of novel viruses [marine thaumarchaea (Swan et al., 2014), and the Microbial Dark Matter project (Rinke et al., 2013)].

Results

The genetic diversity of viral communities in the oxic and anoxic waters of Saanich Inlet was assessed through viral metagenomic data (Table S1) and large-insert fosmids (Table S3). Each fosmid represents a partial genome as it originated from a single strand of viral DNA. Overall, the fosmid sequences lacked similarity to known viral genomes as 5010 of 6459 (77.5%) ORFs across all 34 viral fosmids had no significant homology to viral reference genomes. However, annotation with NCBI's non-redundant database (nr) led to the identification of a putative virus infecting marine thaumarchaea. The viral community composition and genetic content in Saanich Inlet differed between the oxic and anoxic metagenomes from the viral size fraction ($<0.22\,\mu m$) at 10 m ($SI.10_m$) and 200 m ($SI.200_m$), respectively. Collecting viral fosmid and metagenomic sequences from the same samples facilitated direct comparisons of the relative abundance of individual viral types through fragment recruitment of metagenomic reads to identify major contributors to the Saanich Inlet viral assemblages. Additionally, detailed fragment recruitment of viral metagenomic sequences to single-cell genomes (SAGs) revealed prokaryotic genomic regions that are likely from viruses that infect marine thaumarchaea and proteobacteria in the SUP05 clade. Details of these results are presented below.

Saanich Inlet Viral Communities Are Primarily Comprised of Viruses with No Homology to Other Virus Isolate Genomes

Diversity within Viral Metagenomes

Only 16.9% ($SI.10_m$) and 13.1% ($SI.200_m$) of the sequences could be taxonomically assigned based on significant BLASTx hits

to the non-redundant viral genomes in RefSeq (**Figure 1**). The sequences with taxonomic hits were primarily dsDNA viruses from the Order *Caudovirales* (Figures S1, S2) based on GAAS-computed community composition estimates that account for genome length variation among viral taxa (Angly et al., 2009). Within the dsDNA virus fraction, podovirus-like reads were the most abundant ($SI.10_m$: 34.7%, $SI.200_m$: 38.9%), while slightly fewer reads were assigned to siphoviruses ($SI.10_m$: 28.5%, $SI.200_m$: 36.4%); other viruses were 17.7% ($SI.10_m$) and 15.4% ($SI.200_m$), while myovirus-like reads comprised 9% (both $SI.10_m$ and $SI.200_m$) and unclassified viruses in the *Caudovirales*, 5% ($SI.10_m$) and 4% ($SI.200_m$). Viral taxa that each recruited more than 5% of the dsDNA virus reads included *Persicivirga* phage P12024L ($SI.10_m$), *Peligibacter* phage HTVC010P ($SI.10_m$) and *Vibrio* phage pYD21-A ($SI.200_m$). Other dsDNA viruses that each recovered ∼2% of the reads in $SI.10_m$ included *Roseobacter* phage SIO1, *Peligibacter* phage HTVC011P, *Peligibacter* phage HTVC019P, *Celeribacter* phage P12053L, and *Cellulophaga* phage phi10:1 (Figure S1). In $SI.200_m$, ∼2% of the dsDNA reads were assigned to *Peligibacter* phage HTVC010P, *Puniceispirillum* phage HMO-2011, and phages of *Cellulophaga* and *Vibrio* spp (Figure S2). Sequences assigned to phycodnaviruses accounted for only 0.38% ($SI.10_m$) and 0.26% ($SI.200_m$) of the dsDNA reads, and ssDNA viruses totaled 14% ($SI.10_m$) and 10% ($SI.200_m$) of all classified metagenomic sequences.

The estimated richness and overall sequence similarity in the Saanich Inlet viral metagenomes were similar to others from the open ocean, but were higher in $SI.10_m$ than $SI.200_m$ across three nucleotide similarity cutoffs of 75, 90, and 98% when looking at all sequence reads (Figure S3) and when metagenomes were sub-sampled to 50,000 reads (Figure S4). Neither metagenome was sequenced to completion given that the rarefaction curves remained near linear after accounting for all sequences (Figure S3). Re-sampling 50,000 reads per viral metagenome facilitated comparisons across many viral metagenomes obtained with different sequencing efforts. All available marine viral metagenomes were also under-sampled given that ∼48,000 sequence clusters were formed on average per 50,000 sub-sampled reads (Figure S4). At 50,000 sampled reads per viral metagenome and 90% nucleotide similarity, $SI.10_m$ had 48,806 and $SI.200_m$ had 45,980 clusters, and yielded a similar number of sequence clusters as for data from other oceanic viral metagenomes (Figure S4).

Depth and region-based clustering of viral metagenomic data was observed by searching for sequence similarity, despite the effects from under-sampling (Figure S5). Two major clusters were resolved using sequence similarity rather than taxonomic community compositions. One cluster included only oxic viral metagenomes, while the second was comprised of viral metagenomes from below the chlorophyll maximum or from anoxic or low oxygen zones of the water column. The surface or oxic cluster also contained sub-clusters by geographic region (i.e., Line P in the northeast subarctic Pacific Ocean, Line 67 off the coast of Monterey CA, Scripps Pier in San Diego CA, etc…). $SI.10_m$ grouped with other surface ocean viral metagenomes, but not within any of the regional clusters. The low-oxygen or anoxic viral metagenomes were less structured by region than the surface ocean or oxic metagenomes. $SI.200_m$ fell outside of either cluster, making it distinct from viral metagenomes from across the Pacific Ocean. A third major cluster included viral metagenomes from samples pooled from several depths [Arctic_Vir, Gulf_of_Mexico, and British_Columbia, (Angly et al., 2006)], viral metagenomes with shorter sequence lengths (≤250 bp), and one viral metagenome with notable non-marine inputs (Coral_Atoll_Kiritimati, Dinsdale et al., 2008). Although read length may have contributed to the clustering pattern, some of the viral metagenomes with shorter read lengths (<250 bp) also appeared in the surface-oxic cluster. In general, viral communities from the surface ocean were distinguishable from those at depth or with lower oxygen concentrations.

Diversity of Viral Fosmids

From the few recognizable sequences, the taxonomic assignment or classification of each fosmid can offer insights into the virus' lifestyle and possible hosts. Similar to the metagenomes, the limited number of ORFs in the viral fosmids with homology to a viral genome in the RefSeq dataset were primarily similar to members of the Order *Caudovirales* by the best BLASTx hit of each ORF and the last common affiliation (consensus) of all BLASTx hits recovered per fosmid (**Figure 2,** Table S3). The fosmids had significant sequence similarity to several known marine viruses, including pelagiphages, cyanophages, and phages of *Celluphaga* and *Puniceisprillum.* Only five of 34 fosmids, Oxic1_4, Oxic1_9, Oxic1_11, Oxic3_4, and Anoxic3_6, had more than 50% of its ORFs annotated by BLASTx similarity to a protein previously recovered from a viral genome. Three of these fosmids, Oxic1_9, Oxic1_11, and Anoxic3_6, had several ORFs in common with the *Pelagibacter* (SAR11) phage HTVC010P (Figure S6), with Anoxic3_6 and Oxic 1_9 being most similar (52.3% pairwise nucleotide identity) despite an unaligned gap near the putative tail fiber ORFs. Fosmid Oxic1_4 was most similar to another *Pelagibacter* phage HTVC011P and Oxic3_4 was found similar to several *Synechococcus* phage genomes (Syn5, P60, S-CBP42, S-SSM4). Other notable assignments included viruses infecting the genera *Rhizobium, Streptococcus, Vibrio, Enterobacteria* and *Dunaliella,* although confidence in these assignments was limited due to the lack of consistent taxonomy within a fosmid and low amino-acid similarities. Gene assignments by MetaVir were consistent by gene name with manual annotations by BLASTx to the non-redundant database "nr"; taxonomic affiliations from "nr," however, skewed toward prophage regions in cellular organisms and viral fosmid sequences from the Mediterranean Sea (Mizuno et al., 2013b). In summary, annotation of the Saanich Inlet fosmid sequences against reference viral genomes indicated the presence of pelagiphage-like viruses, cyanophages, and many unassigned viruses.

Fragment Recruitment of Metagenomic Reads Indicates Distribution by Depth

Saanich Inlet fosmids recruited more viral metagenomic sequences than the viral genomes in RefSeq or viral fosmids from the Mediterranean Sea (**Figure 3**). Only 0.17% ($SI.10_m$) and 0.05% ($SI.200_m$) of the metagenomic reads had significant sequence similarity to any of the 5580 viral genomes in RefSeq, using a more stringent cutoff of 90% nucleotide identity,

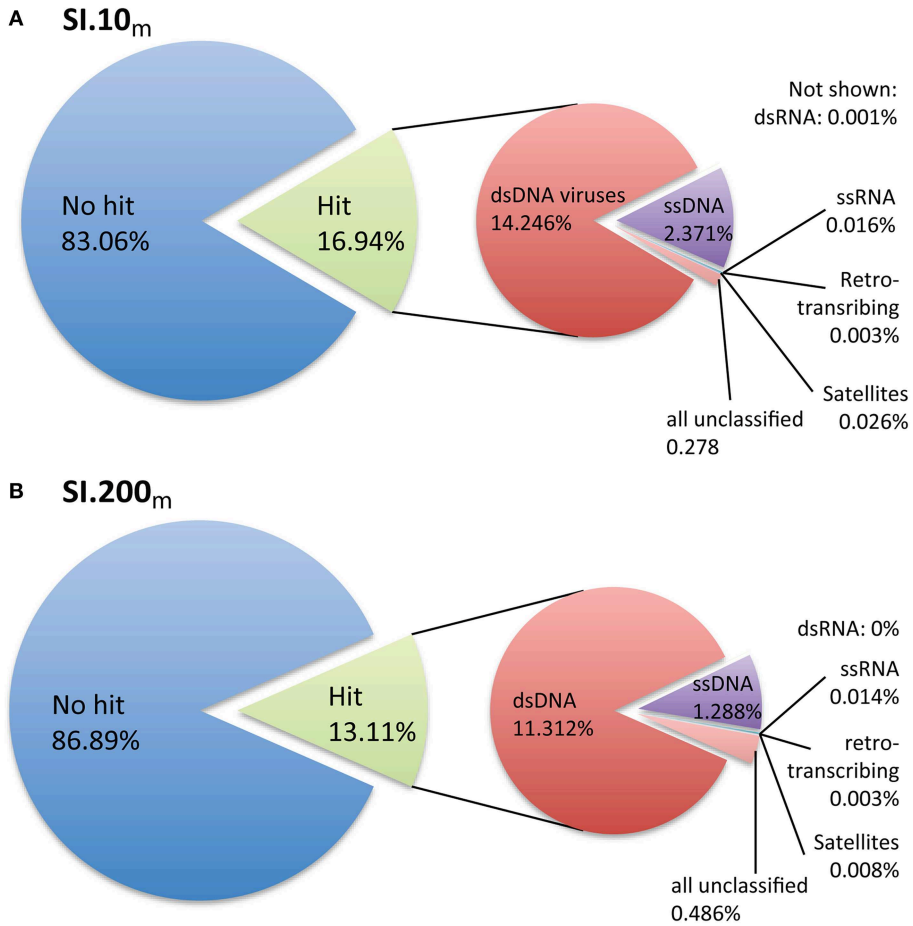


FIGURE 1 | Taxonomic assignment of metagenomic reads based on viral reference genomes. (A) $SI.10_m$ sequences are from oxic waters at 10 m ($n = 2.052,047$) and **(B)** $SI.200_m$ sequences are from waters that are generally anoxic at 200 m ($n = 1.728,968$). Percentages listed are the percent of sequences classified from all metagenomic data per sample.

maximum *e*-value of 10^{-5} and minimum alignment length of 50 bp. Mediterranean Sea viral fosmids (Mizuno et al., 2013a,b) recruited 0.25 and 0.04% of sequences from $SI.10_m$ and $SI.200_m$, respectively. Lastly, all Saanich Inlet viral fosmids collectively recruited 0.78 and 3.78% of sequences from $SI.10_m$ and $SI.200_m$, respectively. Fosmids from 10 m ($SI.10_f$) recruited 0.6% of reads from $SI.10_m$ while fosmids from 200 m ($SI.200_f$) recruited only 0.15% from $SI.10_m$. Conversely, $SI.200_f$ recruited more reads from $SI.200_m$ (2.89%) than $SI.10_m$ (0.75%). Viral fosmids identified from the cellular fraction (SI.Prokaryotic) recruited an additional 0.03% of reads from $SI.10_m$ and 0.14% of reads from $SI.200_m$. $SI.10_f$ also recruited the most reads from the Pacific Ocean Virome Line P (POV.LineP) viral metagenomes, which included viral metagenomes from both surface and deep waters.

Read recruitment from the Saanich Inlet viral metagenomes was unevenly distributed among fosmids (**Figure 4**). Four fosmids (2 from $SI.10_f$, 2 from $SI.200_f$) recruited more than 50 reads per kb (fosmid length) per Gbp (metagenome) from $SI.10_m$. Eight fosmids (1 from $SI.10_f$, 6 from $SI.200_f$, 1 from SI.Prokaryotic) recruited more than 50 reads per kb (fosmid length) per Gbp (metagenome) from $SI.200_m$. Additionally, the four fosmids that recruited the most reads from POV.LineP were all pelagiphage-like (**Figure 4**) although the number of reads recruited by each fosmid differed. Anoxic3_6 and Oxic1_11 recruited over 1500 reads from $SI.10_m$ (77.7 and 100.9 reads per kb per Gbp, respectively) compared to 89 reads and 307 reads (5.1 and 20.6 reads per kb per Gbp, respectively) from $SI.200_m$. In contrast, Oxic1_9 recruited 579 and 112 reads from $SI.10_m$and $SI.200_m$ for 28.8 and 6.3 reads per kb per Gbp, respectively. The fosmids from 10 m, in general, recruited more reads from $SI.10_m$ than $SI.200_m$ and fosmids from 200 m recruited more reads from $SI.200_m$ than $SI.10_m$.

When compared to viral metagenomic data from many different sources, Saanich Inlet fosmids were more similar to viral sequences from marine rather than non-marine sampling locations (Table S2). Specifically, the fosmids primarily recruited sequences from samples in the Moore Marine Phage/Virus Metagenomes project (CAM_PROJ_BroadPhage), which contains viral metagenomic data from throughout the world's oceans.

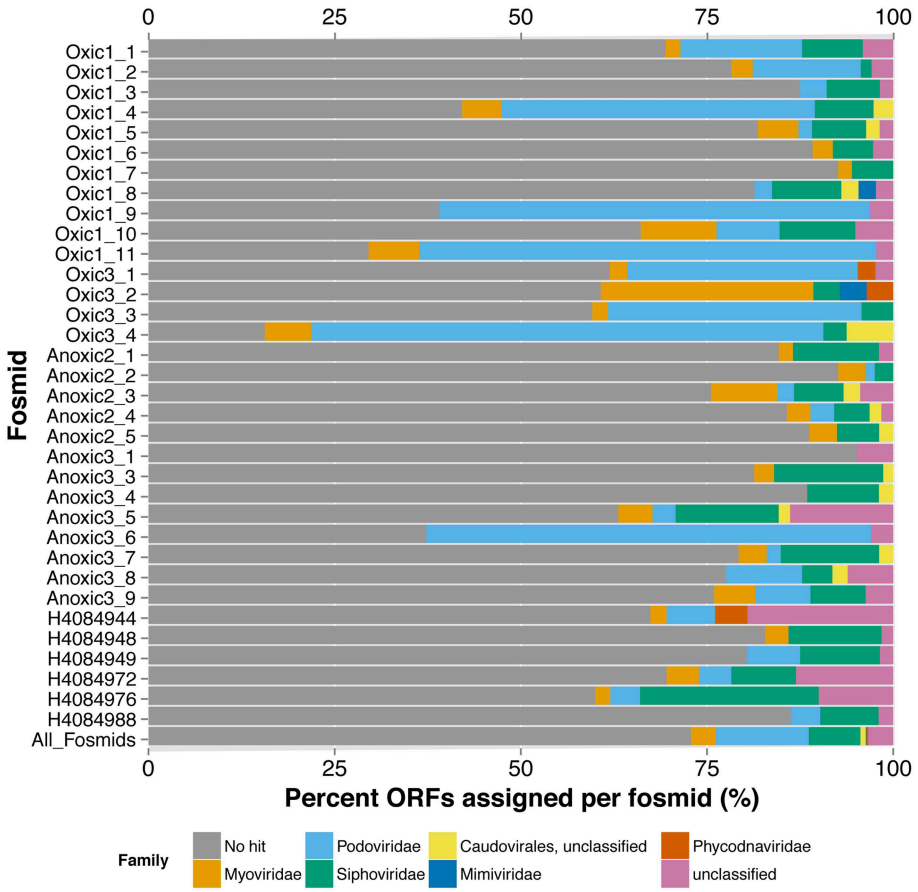


FIGURE 2 | Taxonomic Annotation of Saanich Inlet Fosmids. Classifications are shown per fosmid (each bar) by percent of ORFs assigned to each viral family (x-axis). Bars are colored by classification at family level.

The fosmids with the most metagenomic hits across all of these metagenomes were Oxic1_6, Oxic1_8, Anoxic2_1, Oxic1_1, and Anoxic2_3. Three pelagiphage-like fosmids, Oxic1_9, Oxic1_10, Oxic1_11, recruited reads from nine different metagenome projects, indicating these fosmid sequences originated from viruses widespread in the environment.

Virus Discovery by Paired Analysis of "Omic" Datasets

Leveraging the Non-Redundant "nr" NCBI Database Uncovered Genomic Evidence for Putative Marine Thaumarchaeal Viruses

Manual annotation of the fosmids against the non-redundant "nr" reference database (NCBI) revealed a putative host for fosmid Oxic1_7, the most well-represented fosmid in both Saanich Inlet viral metagenomes. Viral metagenomic reads were recruited across all of Oxic 1_7 except between 35 and 38 kb (7467 reads from SI.10$_m$ and 13,696 reads from SI.200$_m$; **Figure 5A**). Four of 193 ORFs had sequence similarity to siphoviruses, including one hit to the archaeal BJ1 virus when querying the viral reference genomes alone (MetaVir). When the fosmid ORFs were queried against "nr," 25 of 193 ORFs had significant hits (**Figure 5B**). Seven of these ORFs matched the putative thaumarchaeal provirus, Pro-Nvie1, that occurs in the genome of the ammonia-oxidizing thaumarchaeon Candidatus *Nitrososphaera viennensis* strain EN76 isolated from soil (Krupovic et al., 2011). These ORFs included hallmark viral sequences that putatively encode for: terL (terminase, large subunit), protease/major capsid proteins, and tail proteins with an average 30% amino acid identity. Other Oxic1_7 ORFs were similar to DNA methylases and helicases found in other archaea (average 53% amino acid identity across four ORFs) and bacteria (average 50% across 11 ORFs to *Firmicutes*). The three remaining ORFs were similar to hypothetical proteins found in *Batrachochytrium dendrobatidis* ($n = 1$) and EBPR siphovirus 2 ($n = 2$).

Identifying Regions of Possible Viral Origin within Single-Cell Genomes

Putative viral regions in thaumarchaeal and SUP05 SAGs were identified and confirmed by recruitment of viral metagenomic sequences to contigs within the SAGs (**Figure 6**). The two examples detailed below are for host organisms for which little is known about possible host-virus interactions in the ocean. SI.10$_m$ and SI.200$_m$ were also recruited against the "Microbial Dark

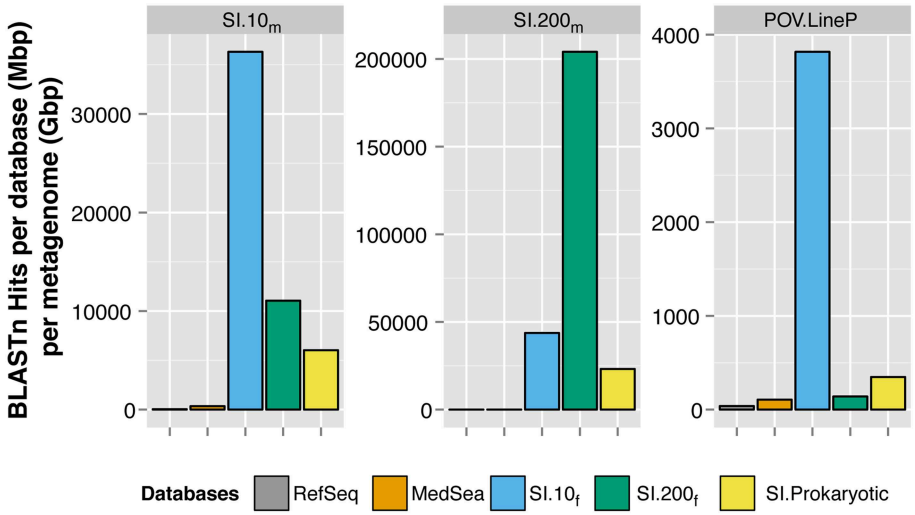


FIGURE 3 | Recruitment of metagenomic sequences to viral reference genomes and fosmids. Each panel represents a BLASTn search of reads from SI.10$_m$ (left), SI.200$_m$ (center), and LineP viromes (POV.LineP) against viral genomes in RefSeq, Mediterranean Sea (MedSea) viral fosmids and Saanich Inlet viral fosmids (SI.10$_f$, SI.200$_f$, SI.Prokaryotic). The cumulative number of hits returned per total bp in each fosmid or genome collection (Mbp) per gigabasepairs (Gbp) of metagenomic sequence data (y-axis) is shown. Note y-axis is different for each panel.

Matter" SAGs from Rinke et al. (2013), but no single contig recruited a significant number of sequences that were distributed somewhat evenly across a region equivalent to a few viral genes.

By recruiting viral metagenomic reads, putative viral regions on contigs reported from SUP05 SAGs were independently verified, particularly with sequences from SI.200$_m$. The same viral contigs and regions were identified as putative viral sequences from marine bacteria SUP05 through the presence of hallmark viral genes (Roux et al., 2014a). Two SUP05 SAGs, in particular, recruited up to 15,643 reads per contig from SI.200$_m$. The first SUP05 SAG (AB.754.J03AB.906) recruited 22,369 reads from SI.200$_m$ and 303 reads from SI.10$_m$ to 6 of its 43 contigs. A second SAG (AB750C22AB.904) recruited 6894 sequences from SI.200$_m$ and 71 reads from SI.10$_m$ to 7 of its 63 contigs. The highest recruiting contig from a SUP05 SAG (AAA160.G15) from SI.10$_m$ recovered only 831 reads, which was not surprising given that SUP05 is typically observed in anoxic waters (Walsh et al., 2009; Wright et al., 2012). This analysis confirmed the presence of these viruses or their close relatives as members of the viral assemblage (or organisms smaller than 0.2 um size fraction) in the anoxic zone of Saanich Inlet.

The same approach was followed to identify putative viral regions in marine thaumarchaea SAGs (**Figure 6**). One thaumarchaeal SAG (AAA288-I14) from Station ALOHA recruited metagenomic reads from both 10 m and 200 m in Saanich Inlet across contigs 23 and 45. Contig 23 recruited 1199 reads from SI.10$_m$ and 246 reads from SI.200$_m$(total = 1445 reads) and contig 45 recruited 371 reads from SI.10$_m$ and 63 reads from SI.200$_m$ (total = 444 reads). Average coverage was 15.4-fold (SI.10$_m$) and 3.4-fold (SI.200$_m$) for contig 23 and 13.5-fold (SI.10$_m$) and 2.6-fold (SI.200$_m$) for contig 45. The two SAG contigs included ORFs which encode for a putative phage tail fiber and other hypothetical proteins found in marine phage genomes (Swan et al., 2014). Viral metagenomic read recruitment to an additional 26 archaeal SAGs from Saanich Inlet resulted in recruitment of 10 or fewer reads each. Thus, these additional SAGs either did not encode genetic content similar to sequences captured in the viral metagenomes or lacked viral regions altogether due to incomplete genome sequencing or natural variation.

Discussion

Advances in nucleic-acid technologies have led to huge increases in viral sequence data; yet, most of these environmental sequences are orphans without a genomic context. Finding a genomic home for these data and ultimately elucidating a function for this viral "dark matter" requires representative virus reference genomes, which can be used to recruit viral metagenomic data. Viral reference genomes may originate from cultured isolates, but with few exceptions, the lack of representative host strains in culture and the enormous microbial diversity in nature means that it is untenable to bring most of the representative cellular diversity into culture. Thus, the vast majority of marine viral reference genomes will not be acquired using culture-based approaches. Moreover, the vast viral sequence diversity in aquatic systems and the relatively short reads provided by current high-throughput sequencing technologies makes it intractable to confidently assemble complete genomes from metagenomes except for RNA (e.g., Culley et al., 2006, 2014) and ssDNA viruses (Tucker et al., 2010; Labonté and Suttle, 2013a,b), which have very small genomes. For dsDNA viruses, reference genomes need to be derived from sequencing large fragments of viral DNA, such as are captured by fosmid cloning (e.g., Garcia-Heredia et al., 2012), targeted metagenomics (e.g., Martinez-Martinez et al., 2014), or potentially single-virus genomes (Allen et al., 2011). Viral reference genomes can provide templates against which

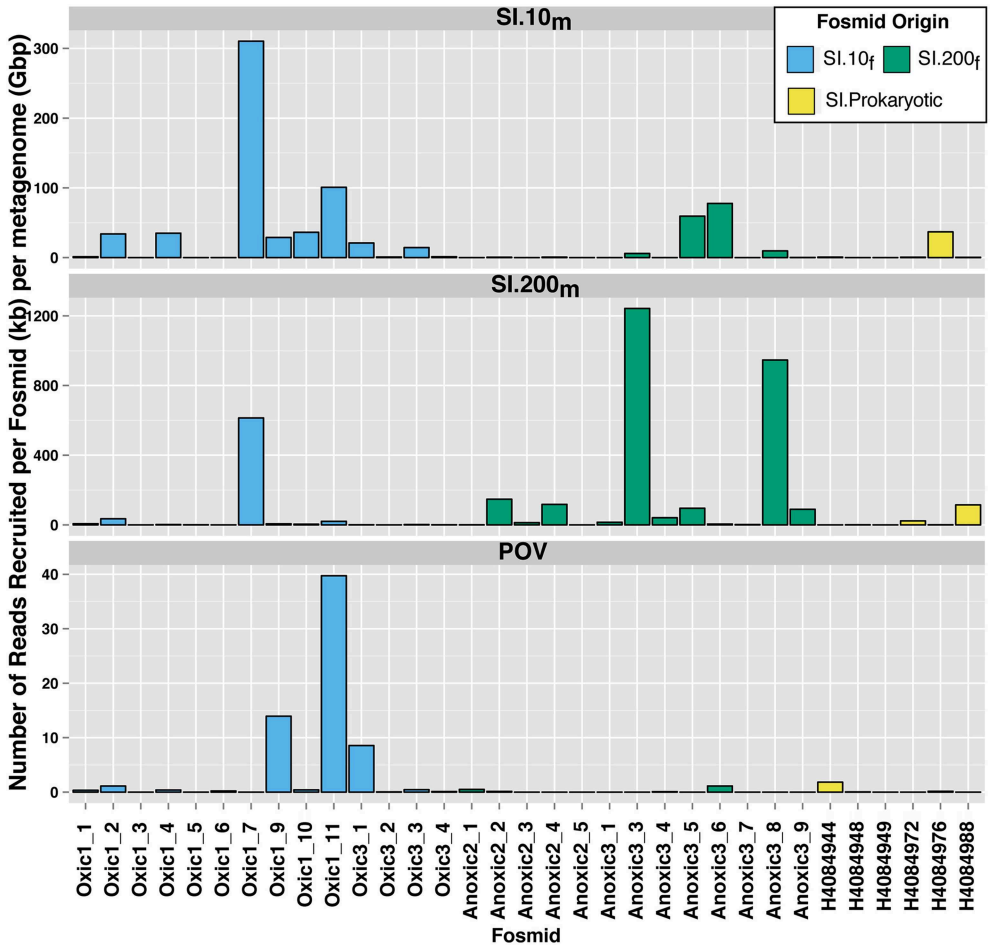


FIGURE 4 | Differential recruitment of metagenomic reads to Saanich Inlet fosmids. The number of hits per fosmid was normalized as the number of BLASTn hits per kb of fosmid length per gigabasepairs of metagenomic sequence data (y-axis).

metagenomic data can be recruited and placed in a genomic context, which will facilitate closing some of the gaps in our knowledge regarding marine virus genomes; these are both essential steps forward to understanding the ecology of marine viruses (Culley, 2013; Bibby, 2014). Our study contributes 34 new partial viral reference genomes from fosmid cloning and sequencing, and identifies genomic fragments of putative viruses infecting marine thaumarchaea and SUP05 proteobacteria. This study also demonstrates that relying on annotation using genomes from cultured viruses alone is a barrier to the discovery of new virus taxa. Moreover, the repeatable recovery of highly similar sequences from both the virus and cellular size fractions of seawater indicates that these sequences likely represent active and common viruses within the marine environment that should be targeted for further investigation.

Combining nucleic-acid sequencing technologies (metagenomic, fosmid, and single-cell genomic datasets) to explore viral diversity and virus-host interactions allowed orphaned metagenomic data from Saanich Inlet to be placed into a genomic context and showed that Saanich Inlet viral communities are distinguishable from those in other environments. Recruitment of metagenomic reads to SAGs highlighted genomic islands of likely viral origin. Comparative analyses between viral fosmid sequences and SAGs uncovered previously unknown viruses and host-virus relationships, such as the putative thaumarchaeal virus sequence from fosmid Oxic1_7. In particular, data from the oxygen-minimum zone provided strong evidence for the presence of these putative viruses infecting marine thaumarchaea and SUP05 proteobacteria, emphasizing that viruses in these environments are relatively understudied. These results show the power of combining environmental genomic approaches to illuminate viral "dark matter" and are discussed in detail below.

Taxonomic Identification of Viral Metagenomic Sequences in Saanich Inlet Was Limited

Prokaryotic communities differ between anoxic waters found at depth and oxic surface waters; thus viral communities would also be expected to differ (e.g., Cassman et al., 2012). However, the scenario in Saanich Inlet is more complex, as stratification is perturbed by deep water renewal shoaling anoxic/sulfidic bottom waters upwards with concomitant changes in microbial community composition (Zaikova et al., 2010; Wright et al.,

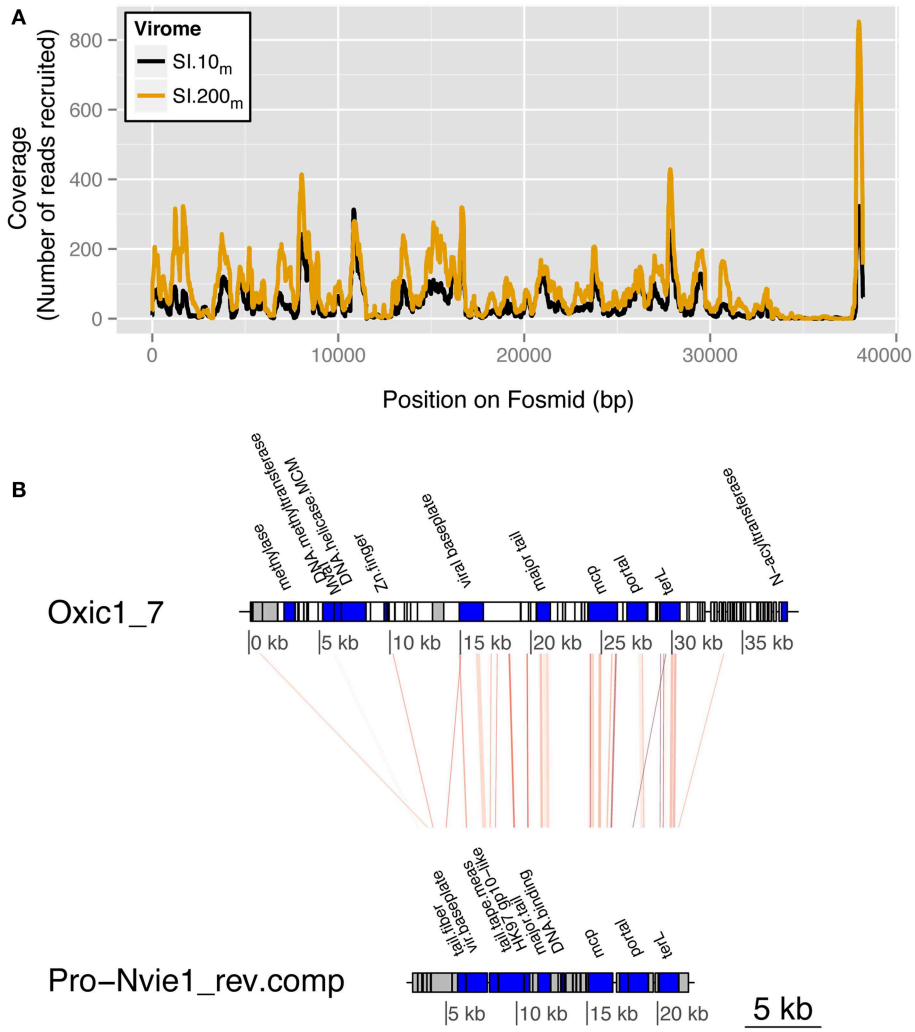


FIGURE 5 | Saanich Inlet fosmid Oxic 1_7 is a putative thaumarchaeal virus. (A) Fragment recruitment coverage (y-axis) by position (bp) along the fosmid (x-axis) is shown for SI.10_m (black) and SI.200_m (gold) for reads with greater than 90% nucleotide identity. **(B)** Regions of genetic similarity by tBLASTx hits to thaumarchaeal provirus Nvie1 (reverse complement) with an *e*-value cutoff of 10^{-5}. The red lines connect the regions of genetic similarity as determined by tBLASTx between the fosmid Oxic1_7 and the provirus Nvie1. ORF blocks are shaded to indicate annotated genes (blue), hypothetical proteins (gray), or no hit in a database (white).

2012). As samples from the same depth were pooled across time, this may be one reason that Saanich Inlet virus communities were not as easily distinguished from those at other marine locations when looking at family- and order-level taxonomic classifications. Metagenomic sequences were assigned to many taxa of marine viruses when queried against RefSeq, a database of viral reference genomes (**Figure 1**, Figures S1, S2). As this reference database is dominated by sequences from viruses within the *Caudovirales*, this in turn dictated the taxonomic placement of the metagenomic reads. For both metagenomes, ∼85% of the metagenomic reads were not assigned to a taxonomy (**Figure 1**). In fact, the percent of metagenomic reads classified per sample ranged from 0.1 to 28.5% (average 15.2%) across all marine viral metagenomic datasets (e.g., Line P, Line 67, Scripps Pier) used for comparative analyses. Although SI metagenomic samples were not sequenced to completion, the diversity estimates were similar to those obtained from other ocean sites (Figure S4). What is evident from the data is that improved classification and ecological interpretation of viral metagenomic data requires more representative viral genomes in the reference databases.

Comparisons based on taxonomic classification may obscure differences among samples as most sequences remain unassigned, but clustering by sequence similarity resolved differences in viral metagenomic data from oxic and anoxic waters (Figure S5). Saanich Inlet sequences grouped with POV metagenomic data according to depth, consistent with oxygen concentrations in the two environments; this is one of the few comparisons to demonstrate clustering of viral metagenomic data by ecological niche (Hurwitz et al., 2015). These results collectively suggest

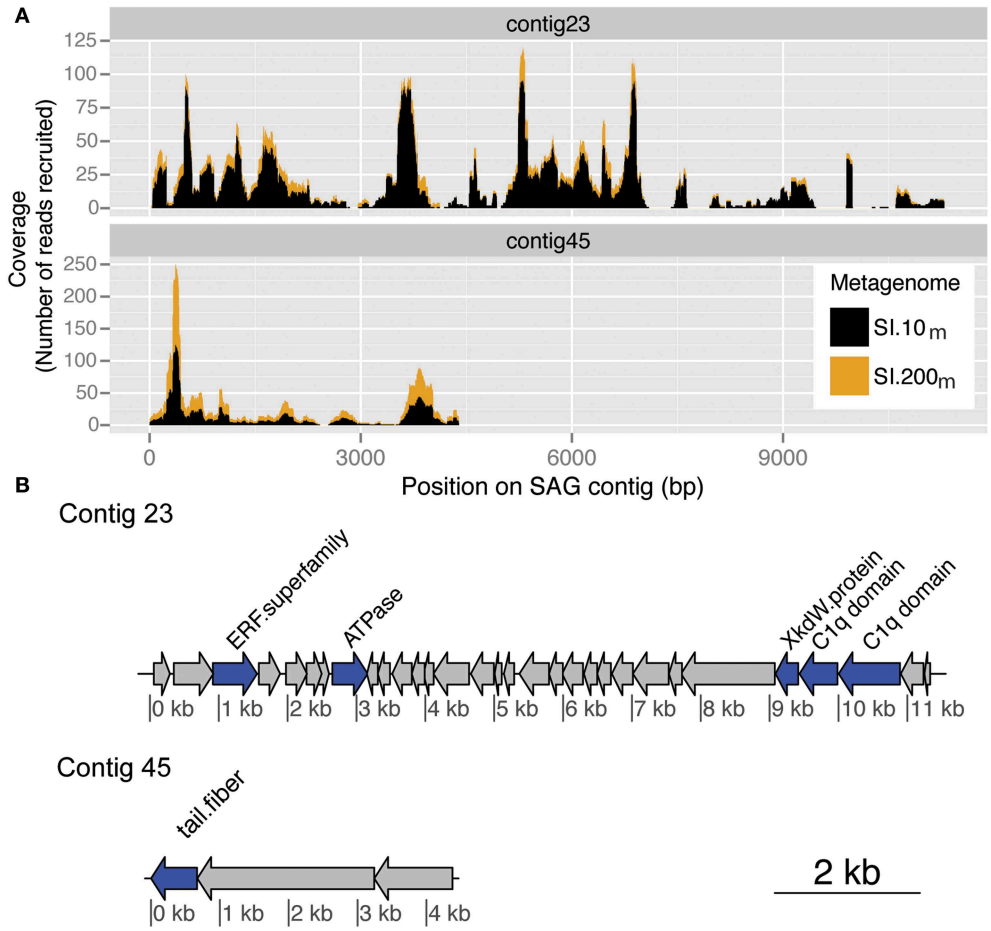


FIGURE 6 | Putative viral contigs recovered from a marine thaumarchaeal SAG. (A) Coverage by metagenomic reads from SI.10$_m$ (black) and SI.200$_m$ (gold) with greater than 90% nucleotide similarity to a marine thaumarchaeal SAG for contig 23 and contig 45. Coverage is cumulative (stacked height of histogram) across both metagenomes. **(B)** ORF annotations for each SAG contig as described in the Integrated Microbial Genomes (IMG) system. ORF color indicates whether it is a named (blue) or hypothetical (gray) protein.

that the composition of viral communities is predictable based on abiotic or biotic influences in the local environment.

The taxonomic identities assigned to the metagenomic sequences also included many ssDNA viruses. Although several reference genomes from ssDNA viruses have been assembled from other Saanich Inlet metagenomic data (Labonté and Suttle, 2013a,b), the percent contribution of metagenomic sequences belonging to ssDNA viruses in this study is likely over-estimated due to the biases associated with multiple displacement amplification (Polson et al., 2010; Kim and Bae, 2011). For this reason, our analysis focused on dsDNA viral communities and the novel diversity recovered in the metagenomic data in relation to the viral fosmid and single-cell data from this study and others.

Viral Diversity Recovered by Fosmids as Genome Proxies

Fosmid cloning and sequencing has been used for recovering complete and partial viral genomes from seawater (Mizuno et al., 2013a,b)and hypersaline environments (Garcia-Heredia et al., 2012). Although it is low throughput and time-intensive, fosmid cloning captures up to ~40 kb of DNA from a single virus; whereas metagenomic sequences lack a genomic context. Both methods, however, yield viral genomic data without the need for culturing. Fosmid cloning may also facilitate recovery of uncommon taxa due to bias for sequences with higher G+C content (Danhorn et al., 2012).

Isolate-based viral genomes provided excellent templates against which closely related fosmids could be compared. As proof of principle, the genomes of pelagiphages, which are common in the marine environment and abundant in most viral metagenomic data (Zhao et al., 2013), were compared to three SI fosmids that had multiple BLASTx hits to pelagiphage ORFs (Figure S6). The comparisons confirmed that based on genomic content and organization the fosmids contained DNA from close relatives of pelagiphage isolates, although there was evidence of population or strain differences between the isolates and the viruses represented by the fosmids. Annotation of the remaining fosmids using traditional reference databases provided clues

as to the taxonomic classification of each fosmid and their potential hosts (**Figure 2**, Table S3). However, these results were often inconclusive due to different taxonomic assignments by BLASTx similarity to ORFs within a single fosmid and so demonstrate the novelty of the representative viruses captured by the fosmid sequences.

Virus Ecology Inferred from Paired Analysis of Molecular Datasets

More than 1000 fosmids covering a spectrum of viral taxa have been sequenced from the Mediterranean Sea (Mizuno et al., 2013a,b; Rodriguez-Valera et al., 2014), yet the 34 viral fosmids sequenced from Saanich Inlet in this study recruited more metagenomic sequences than all of the MedSea fosmids combined. Thus, a larger proportion of metagenomic sequences from Saanich Inlet could be assigned to a genomic context using locally derived fosmid sequences than could be assigned to a much larger database of fosmid sequences from another location (**Figure 3**). There was minimal overlap in the metagenomic sequences from each database that were recruited to the fosmids; the SI fosmids tended to have higher sequence similarity to the metagenomic sequences than did the MedSea fosmids for the few that did overlap (data not shown). Given the environmental differences between the locations and sampling depths, these results are not surprising but show that viral communities are specific to their environment, consistent with the metagenomic sequence similarity clustering patterns (Figure S5).

Recruitment of metagenomic sequences from $SI.10_m$ and $SI.200_m$ to the SI fosmids also highlighted depth-dependent distributions as fosmids from 10 m recruited more sequences from $SI.10_m$ than $SI.200_m$ and vice versa (**Figures 3**, **4**). These differences persisted against a backdrop of sample pooling, DNA amplification during sample preparation, and seasonal dynamics that would be expected to mask these differences. In particular, Anoxic3_3 and Oxic 1_7 represent key members of the Saanich Inlet viral assemblage during our study. Several hundreds to thousands of metagenomic sequences shared similarity with these two fosmids that represent two distinct viral genomes (**Figure 4**). As such, they may be excellent targets for developing PCR primer sets that could be used to track fluctuations in virus populations over time and depth or to query single-cell genomes to identify possible hosts.

Virus Discovery through Fosmids as Genome Proxies

Evidence for marine archaeal viruses can be inferred from the similarity of sequences recovered from the free virus size fraction ($<0.22\,\mu m$) in Saanich Inlet to a putative thaumarchaeal provirus (this study), CRISPR regions in a thaumarchaeal genome (Spang et al., 2012) and in metagenomic data from a hypersaline lake (Emerson et al., 2013), and detection of genomic islands in single-cell genomes (Swan et al., 2014). Manual annotation with "nr" provided a consistent taxonomic assignment for fosmid Oxic1_7 as a plausible virus of marine thaumarchaea, an identity that would have been missed by annotation using viral reference genomes alone (**Figure 5**). The identification stems from significant sequence similarity to a putative provirus genome, Pro-Nvie1, recovered from a soil archaeal genome. Highly similar metagenomic sequences with greater than 90% nucleotide identity to Oxic1_7 were recovered from both sampled depths, confirming the presence of a closely related thaumarchaea virus in the viral size fraction at Saanich Inlet. The recovery of several ORFs within Oxic1_7 with similarity to the same provirus and the hyperthermophilic archaeal virus BJ1 implies that the DNA from this fosmid originated from an archaeal virus.

Viral sequences occurred in single-cell genomic data from bacteria belong to the SUP05 clade (Roux et al., 2014a), which are abundant in the anoxic waters of Saanich Inlet (Walsh et al., 2009), and in metagenomic data from seawater and vent fluid from the Lau Basin (Anantharaman et al., 2014). Highly similar sequences to $SI.200_m$ were found in the SUP05 single-cell genomic data, but not in metagenomic data from Lau Basin, likely due to the many environmental differences between the sites. Many sequences from $SI.200_m$ were also similar to contigs from thaumarchaeal SAGs from the Pacific Ocean (Swan et al., 2014), that also contained putative viral genes (**Figure 6**).

Marine thaumarchaeal viruses as active players in the ocean, particularly in OMZs, would directly affect biogeochemical cycling. The viral shunt is most often viewed in the context of viral-driven recycling of carbon, nitrogen, and phosphorus (Fuhrman, 1999; Wilhelm and Suttle, 1999; Shelford et al., 2012; Weitz and Wilhelm, 2012). In addition to being key nitrifiers in the ocean (Francis et al., 2007), marine thaumarchaea have also been implicated as important remineralizers of cobalamin (vitamin B12) in the ocean (Doxey et al., 2015). Many cellular processes require cobalamin as an enzymatic cofactor and it is often a limiting factor in cellular growth (Sañudo-Wilhelmy et al., 2014). If highly active, viral infection of marine thaumarchaea and SUP05 would have significant potential effects on the nitrogen, sulfur and other biogeochemical cycles that are unaccounted for in the current flux budgets.

Concluding Remarks

Viral ecology has benefited greatly from the adoption of nucleic-acid technologies to assess viral diversity and coding potential. Higher-throughput sequencing, lower costs, and new methods to recover, amplify, and target viral particles and nucleic acids are continuing to push research in new directions. Although assessment of the taxonomic composition of the Saanich Inlet viral community was limited by the availability of reference genomes, the 34 new fosmid sequences obtained in this study provided a genomic context for a significant and otherwise orphaned proportion of the viral metagenomic data. Taken together, these results highlight the power of combining sequencing approaches and the resulting data to interrogate viral diversity and discover potential virus-host interactions. For example, our analysis of viral metagenomic, fosmid sequences, and prokaryotic single-cell genomes together provided genetic evidence for a likely active and common presence of viruses infecting thaumarchaea in the global ocean. These findings may have been less convincing if each dataset was only considered on its own.

Additionally, this study is one of only a handful of studies completed to date that demonstrates clustering of viral communities by ecological niche using metagenomic data. Strategic sampling and genetic exploration of under-explored areas, such as anoxic waters, will provide important resources for understanding not only the genetic diversity and genetic potential of marine viruses but also their contributions to nutrient cycling and ecosystem services.

Author Contributions

DMW and CAS conceived the study. DMW and RAW carried out the laboratory work and CTC, DMW, and RAW conducted the bioinformatics analyses. CTC and DMW wrote the paper with input and revisions from CAS. RAW and SJH contributed to discussions of the results and article content. All authors have reviewed and agreed to the final content.

Acknowledgments

We thank the many members of the Suttle and Hallam labs for collecting and processing the samples and the captain and crew aboard the MSV John Strickland for logistical support. Specific thanks to Caroline Chenard, Jessica Labonte, Tyler Nelson, and Alyse Hawley for assistance with method development and helpful discussions, and Amy Chan for technical support. We would also like to thank Jan Finke, Julia Gustavsen, and Emma Shelford for comments that improved the manuscript. This research was supported by the Tula Foundation and grants awarded to CAS and SJH including NSERC Discovery, Canada Foundation for Innovation (CFI), and the Canadian Institute for Advanced Research (CIFAR). Sample collection was facilitated through ship-time grants from NSERC awarded to PD Tortell and SJ Hallam. Access to sequencing was funded by the Gordon and Betty Moore Foundation through GBMF1799 to the Broad Institute, and by NSERC and the Tula Foundation for using facilities at the Génome Québec Innovation Centre at McGill University.

Supplementary Material

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fmicb.2015.00265/abstract

References

Allen, L. Z., Ishoey, T., Novotny, M. A., McLean, J. S., Lasken, R. S., and Williamson, S. J. (2011). Single virus genomics: a new tool for virus discovery. *PLoS ONE* 6:e17722. doi: 10.1371/journal.pone.0017722

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., and Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. *Science* 344, 757–760. doi: 10.1126/science.1252229

Anderson, J., and Devol, A. (1973). Deep water renewal in Saanich Inlet, an intermittently anoxic basin. *Estuar. Coast. Mar. Sci.* 1, 1–10. doi: 10.1016/0302-3524(73)90052-2

Andersson, A. F., and Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320, 1047–1050. doi: 10.1126/science.1157358

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C. A., et al. (2006). The marine viromes of four oceanic regions. *PLoS Biol.* 4:e368. doi: 10.1371/journal.pbio.0040368

Angly, F. E., Willner, D., Prieto-Davó, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., et al. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol.* 5:e1000593. doi: 10.1371/journal.pcbi.1000593

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9:75. doi: 10.1186/1471-2164-9-75

Bibby, K. (2014). Improved bacteriophage genome data is necessary for integrating viral and bacterial ecology. *Microb. Ecol.* 67, 242–244. doi: 10.1007/s00248-013-0325-x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J. M., Segall, A. M., Mead, D., et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14250–14255. doi: 10.1073/pnas.202488399

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. doi: 10.1038/ismej.2008.31

Cassman, N., Prieto-Davó, A., Walsh, K., Silva, G. G. Z., Angly, F. E., Akhter, S., et al. (2012). Oxygen minimum zones harbour novel viral communities with low diversity. *Environ. Microbiol.* 14, 3043–3065. doi: 10.1111/j.1462-2920.2012.02891.x

Culley, A. I. (2013). Insight into the unknown marine virus majority. *Proc. Natl. Acad. Sci. U.S.A.* 110, 12166–12167. doi: 10.1073/pnas.1310671110

Culley, A. I., Lang, A. S., and Suttle, C. A. (2006). Metagenomic analysis of coastal RNA virus communities. *Science* 312, 1795–1798. doi: 10.1126/science.1127404

Culley, A. I., Mueller, J. A., Belcaid, M., Wood-Charlson, E. M., Poisson, G., and Steward, G. F. (2014). The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *mBio* 5, e01210–e01214. doi: 10.1128/mBio.01210-14

Danhorn, T., Young, C. R., and DeLong, E. F. (2012). Comparison of large-insert, small-insert and pyrosequencing libraries for metagenomic analysis. *ISME J.* 6, 2056–2066. doi: 10.1038/ismej.2012.35

Darling, A. E., Mau, B., and Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* 5:e11147. doi: 10.1371/journal.pone.0011147

Deng, L., Ignacio-Espinoza, J. C., Gregory, A. C., Poulos, B. T., Weitz, J. S., Hugenholtz, P., et al. (2014). Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. *Nature* 513, 242–245. doi: 10.1038/nature13459

Dinsdale, E. A., Pantos, O., Smriga, S., Edwards, R. A., Angly, F. E., Wegley, L., et al. (2008). Microbial ecology of four coral atolls in the northern Line Islands. *PLoS ONE* 3:e1584. doi: 10.1371/journal.pone.0001584

Doxey, A. C., Kurtz, D. A., Lynch, M. D. J., Sauder, L. A., and Neufeld, J. D. (2015). Aquatic metagenomes implicate *Thaumarchaeota* in global cobalamin production. *ISME J.* 9, 461–471. doi: 10.1038/ismej.2014.142

Emerson, J. B., Andrade, K., Thomas, B. C., Norman, A., Allen, E. E., Heidelberg, K. B., et al. (2013). Virus-Host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* 2013:370871. doi: 10.1155/2013/370871

Francis, C. A., Beman, J. M., and Kuypers, M. M. M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *ISME J.* 1, 19–27. doi: 10.1038/ismej.2007.8

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* 399, 541–548. doi: 10.1038/21119

Garcia-Heredia, I., Martin-Cuadrado, A.-B., Mojica, F. J. M., Santos, F., Mira, A., Antón, J., et al. (2012). Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. *PLoS ONE* 7:e33802. doi: 10.1371/journal.pone.0033802

Guy, L., Kultima, J. R., and Andersson, S. G. E. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 26, 2334–2335. doi: 10.1093/bioinformatics/btq413

Hurwitz, B. L., Brum, J. R., and Sullivan, M. B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core' and "flexible" Pacific Ocean Virome. *ISME J.* 9, 472–484. doi: 10.1038/ismej.2014.143

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. doi: 10.1371/journal.pone.0057355

Kim, K. H., and Bae, J. W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* 77, 7663–7668. doi: 10.1128/AEM.00289-11

Krupovic, M., Spang, A., Gribaldo, S., Forterre, P., and Schleper, C. (2011). A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. *Biochem. Soc. Trans.* 39, 82–88. doi: 10.1042/BST0390082

Labonté, J. M., and Suttle, C. A. (2013a). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Front. Microbiol.* 4:404. doi: 10.3389/fmicb.2013.00404

Labonté, J. M., and Suttle, C. A. (2013b). Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* 7, 2169–2177. doi: 10.1038/ismej.2013.110

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Larsen, A., Flaten, G. A. F., Sandaa, R.-A., Castberg, T., Thyrhaug, R., Erga, S. R., et al. (2004). Spring phytoplankton bloom dynamics in Norwegian coastal waters: microbial community succession and diversity. *Limnol. Oceanogr.* 49, 180–190. doi: 10.4319/lo.2004.49.1.0180

Leplae, R., Lima-Mendez, G., and Toussaint, A. (2009). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res.* 38, D57–D61. doi: 10.1093/nar/gkp938

Magoč, T., and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27, 2957–2963. doi: 10.1093/bioinformatics/btr507

Martinez-Martinez, J., Swan, B. K., and Wilson, W. H. (2014). Marine viruses, a genetic reservoir revealed by targeted viromics. *ISME J.* 8, 1079–1088. doi: 10.1038/ismej.2013.214

Mizuno, C. M., Rodriguez-Valera, F., Garcia-Heredia, I., Martin-Cuadrado, A.-B., and Ghai, R. (2013a). Reconstruction of novel cyanobacterial siphovirus genomes from mediterranean metagenomic fosmids. *Appl. Environ. Microbiol.* 79, 688–695. doi: 10.1128/AEM.02742-12

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., and Ghai, R. (2013b). Expanding the marine virosphere using metagenomics. *PLoS Genet.* 9:e1003987. doi: 10.1371/journal.pgen.1003987

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Polson, S. W., Wilhelm, S. W., and Wommack, K. E. (2010). Unraveling the viral tapestry (from inside the capsid out). *ISME J.* 5, 165–168. doi: 10.1038/ismej.2010.81

Ray, J., Dondrup, M., Modha, S., Steen, I. H., Sandaa, R.-A., and Clokie, M. R. J. (2012). Finding a needle in the virus metagenome haystack - micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS ONE* 7:e34238. doi: 10.1371/journal.pone.0034238

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. doi: 10.1038/nature12352

Rodriguez-Valera, F., Mizuno, C. M., and Ghai, R. (2014). Tales from a thousand and one phages. *Bacteriophage* 4:e28265. doi: 10.4161/bact.28265

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., et al. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* 27, 3074–3075. doi: 10.1093/bioinformatics/btr519

Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., et al. (2014a). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *Elife* 3:e03125. doi: 10.7554/eLife.03125

Roux, S., Tournayre, J., Mahul, A., Debroas, D., and Enault, F. (2014b). Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* 15:76. doi: 10.1186/1471-2105-15-76

Sañudo-Wilhelmy, S. A., Gómez-Consarnau, L., Suffridge, C., and Webb, E. A. (2014). The role of B vitamins in marine biogeochemistry. *Annu. Rev. Mar. Sci.* 6, 339–367. doi: 10.1146/annurev-marine-120710-100912

Shackelton, L. A., and Holmes, E. C. (2004). The evolution of large DNA viruses: combining genomic information of viruses and their hosts. *Trends Microbiol.* 12, 458–465. doi: 10.1016/j.tim.2004.08.005

Sharon, I., Tzahor, S., Williamson, S. J., Shmoish, M., Man-Aharonovich, D., Rusch, D. B., et al. (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME J.* 1, 492–501. doi: 10.1038/ismej.2007.67

Shelford, E. J., Middelboe, M., Moller, E. F., and Suttle, C. A. (2012). Virus-driven nitrogen cycling enhances phytoplankton growth. *Aquat. Microb. Ecol.* 66, 41–46. doi: 10.3354/ame01553

Spang, A., Poehlein, A., Offre, P., Zumbrägel, S., Haider, S., Rychlik, N., et al. (2012). The genome of the ammonia−oxidizing Candidatus *Nitrososphaera gargensis*: insights into metabolic versatility and environmental adaptations. *Environ. Microbiol.* 14, 3122–3145. doi: 10.1111/j.1462-2920.2012.02893.x

Steward, G. F., Montiel, J. L., and Azam, F. (2000). Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol. Oceanogr.* 45, 1697–1706. doi: 10.4319/lo.2000.45.8.1697

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

Suttle, C. A., Chan, A. M., and Cottrell, M. T. (1991). Use of ultrafiltration to isolate viruses from seawater which are pathogens of marine phytoplankton. *Appl. Environ. Microbiol.* 57, 721–726.

Swan, B. K., Chaffin, M. D., Martinez-Garcia, M., Morrison, H. G., Field, E. K., Poulton, N. J., et al. (2014). Genomic and metabolic diversity of marine group I *Thaumarchaeota* in the mesopelagic of two subtropical gyres. *PLoS ONE* 9:e95380. doi: 10.1371/journal.pone.0095380

Tucker, K. P., Parsons, R., Symonds, E. M., and Breitbart, M. (2010). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J.* 5, 822–830. doi: 10.1038/ismej.2010.188

Walsh, D. A., Zaikova, E., Howes, C. G., Song, Y. C., Wright, J. J., Tringe, S. G., et al. (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* 326, 578–582. doi: 10.1126/science.1175309

Weitz, J. S., and Wilhelm, S. W. (2012). Ocean viruses and their effects on microbial communities and biogeochemical cycles. *F1000 Biol. Rep.* 4:17. doi: 10.3410/B4-17

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea. *Bioscience* 49, 781–788. doi: 10.2307/1313569

Williamson, S. J., Allen, L. Z., Lorenzi, H. A., Fadrosh, D. W., Brami, D., Thiagarajan, M., et al. (2012). Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* 7:e42047. doi: 10.1371/journal.pone.0042047

Winget, D. M., Helton, R. R., Williamson, K. E., Bench, S. R., Williamson, S. J., and Wommack, K. E. (2011). Repeating patterns of virioplankton production within an estuarine ecosystem. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11506–11511. doi: 10.1073/pnas.1101907108

Winter, C., Smit, A., Herndl, G. J., and Weinbauer, M. G. (2004). Impact of virioplankton on archaeal and bacterial community richness as assessed in seawater batch cultures. *Appl. Environ. Microbiol.* 70, 804–813. doi: 10.1128/AEM.70.2.804-813.2004

Wright, J. J., Konwar, K. M., and Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. *Nat. Rev. Microbiol.* 10, 381–394. doi: 10.1038/nrmicro2778

Zaikova, E., Walsh, D. A., Stilwell, C. P., Mohn, W. W., Tortell, P. D., and Hallam, S. J. (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ. Microbiol.* 12, 172–191. doi: 10.1111/j.1462-2920.2009.02058.x

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., et al. (2013). Abundant SAR11 viruses in the ocean. *Nature* 494, 357–360. doi: 10.1038/nature11921

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on