Metagenomics-enabled viral ecology to advance our understanding of human and environmental microbiomes

By

Kristopher Kieft

A dissertation submitted in partial fulfillment of

The requirements for the degree of

Doctor of Philosophy

(Microbiology)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: May 24, 2022

The dissertation is approved by the following members of the Final Exam Committee:

Karthik Anantharaman, Assistant Professor, Bacteriology
Katherine McMahon, Professor, Bacteriology, Civil & Environmental Engineering
Nicole Perna, Professor, Genetics
Christy Remucal, Associate Professor, Civil & Environmental Engineering
Emily Stanley, Professor, Zoology

Abstract

Viruses are omnipresent and influence life processes on micro and macro scales. All viruses require a host cell to replicate, which they accomplish by hijacking the host cells' machinery and resources. Due to this, viruses, especially those that infect microorganisms, are keystone controllers of nutrient and biomass recycling, ecosystem-wide biogeochemical cycling, host population abundances, community interactions, evolutionary momentum, and more. Understanding viruses is critical to the advancement of diverse fields such as agriculture, medicine, industry, ecosystem science, oceanography, and biogeochemistry. Despite their importance, viruses remain difficult to study as they cannot be easily cultivated in a laboratory. Cultivation independent approaches such as metagenomics and viromics can enable the study of uncultivated viruses directly from the environment, though there is a lack of accurate bioinformatics tools to analyze viruses from metagenomes.

I am interested in exploring the ecology, evolution, and diversity of viruses in nature and investigating their impacts on global biogeochemical cycles. My dissertation is categorized into two main Aims. Aim one of my research is to use systems biology approaches to expand our understanding of viruses by quantifying their impacts on microorganisms and ecosystems. I will specifically focus on viral encoded genes for assimilatory (Chapter 2) and dissimilatory (Chapter 3) sulfur metabolism. Aim two of my dissertation research is to develop novel bioinformatics tools for the study of viruses in metagenomes and viromes. These approaches will serve as the foundation of next generation virology and studying how viruses fit within microbial communities and global biogeochemical cycles by identifying viral genomes (Chapter 4), deciphering active viruses from DNA sequences (Chapter 5), and constructing viral genomes from mixed communities (Chapter 6). Finally, I provide my observations on where the field of viral genomics

stands, established conventions that would benefit from change, and future prospects of the field as a whole (Chapter 7).

Acknowledgements

Without support from others graduate school is impossible, and I'm immensely grateful for the encouragement and guidance I've received along the way. To everyone not specifically mentioned below, I want to thank you all for helping me with my accomplishments and supporting me during the failures.

Thank you to my advisor Karthik Anantharaman, for the one-on-one chats about science or anything, exceptional mentorship and integrity, interest in seeing me thrive, and everything else that won't fit here. With his guidance I found my research passions that set up my future career and path. Thank you for all the letters of recommendation and the many hours helping me succeed. It was a pleasure to join the lab and I'm excited to see where it grows from here.

Thank you to my lab mates who were truly supportive friends. It was fun to see everyone come on board and help build an enjoyable place to work, hang out and chat. From constantly hearing me talk about viruses and being manuscript co-authors and contributors to meeting up at happy hours, I'll always be grateful for the support and experiences they've all provided.

Thank you to my committee members, Trina McMahon, Nicole Perna, Christy Remucal and Emily Stanley, for selflessly encouraging my success. Your questions and discussions have helped to prop me up and train me to be a better scientist. I left each committee meeting with higher spirits and greater goals.

Thank you to all of the Microbiology Doctoral Training Program, my cohort of friends, and the program coordinators Cathy Davis-Gray and Terra Theim. I couldn't have asked for a better group of friends, peers, and mentors to ensure my PhD experience was enjoyable. The community is filled with people I'll miss.

Thank you to everyone from undergrad at Central Michigan University, the Honors program, my professors, and all the friends I met along the way that helped me get to where I am now. I especially want to thank my undergrad lab advisor Michael Conway and for getting me interested in viruses, supporting my mistakes and successes in the lab, and encouraging me to be a better scientist.

Thank you to all of my friends from Michigan who have been there for as long as I can remember. Science isn't everything, and without this group of people I don't know where I'd be now.

Thank you to my family! To my parents and in-laws for being proud and excited even if microbiology was a different language; to my brother for endless support and scientific insights; to my dog for being a relief from stress; to my wife, Sarah, for EVERYTHING (the love, support, encouragement, proofreading, pulling me away from work, listening to crazy ideas, giving advice, and everything that got me through nearly a decade of college). Without my family none of this would be possible and I cannot thank them enough.

Table of Contents

List of Figures

Chapter Two

Chapter Three

Chapter Four

Chapter Five

List of Tables

Chapter 1: Introduction

The virosphere is intertwined with life

In this dissertation I discuss the study of viruses in human and natural systems. Viruses are intracellular parasites of living cells, or even other viruses, that have the sole objective of replicating their genomes and propagating. At the most basic level, viruses are nothing more than a genome, containing all instructions to replicate, and a proteinaceous capsid to protect the genome. Viral genomes, in contrast to living organisms, can be comprised of ssDNA, dsDNA, (+\-) ssRNA, dsRNA, and hybrids of DNA and RNA[1,2].

Viral particles on Earth, encompassing the virosphere[3,4], outnumber every living cellular organism[5,6]. Most such viruses infect bacteria (bacteriophage, phage) and are omnipresent in all studied environments. Although viruses that infect humans and other animals are conspicuous everywhere we look—a prominent example being the SARS-CoV2 pandemic that began in 2019—phages are often hidden from view or overlooked. Although viruses are not considered living, they are a central component of all life; life on Earth is intertwined with viruses[7–10].

The negative impact viruses have on life can be seen everywhere. As humans, we have become persistently aware of viral morbidity and mortality that pose threats to our health, wellbeing, and economies[11–13]; viral infections of crops, produce, and livestock are a major burden of food industries[14–18]. However, some impacts of viruses are beneficial, yet not so evident. For example, the formation of the placental barrier can be attributed in part to viral genes long ago co-opted by mammals[19,20] and that phages may be co-opted by the human immune system[21]. Modern biotechnology has thrived from biological discoveries owing in part to viruses, such as the utility of CRISPR-Cas gene editing systems[22–24], plasmid expression vectors[25,26], vaccine delivery

systems[27,28], phage therapy treatments[29–31], and more[32–34]. On a broad scale, viruses are drivers of system-wide processes, such as organic and inorganic nutrient turnover, which maintain healthy ecosystems[5,35–38]. In this dissertation, I will explore and provide evidence to the latter point, focusing on how viruses are key players in diverse ecosystem processes on global scales.

Virus infection mechanics

Viruses follow a general infection cycle for replication. At first, a proteinaceous particle containing a viral genome (virion) contacts a susceptible host cell. Through various mechanisms, such as protein-protein receptor binding or protein-lipopolysaccharide binding, the virus physically attaches to the host. Once attached, the viral genome is injected or endocytosed into the host cytoplasm or can be directed to the nucleus. In the case of phages, the viral genome remains within the cytoplasm or specialized replication compartments. At this stage, viruses differ considerably in the mechanism of infection. Some viruses temporarily end their infection here by integrating or stably maintaining their genome within the host (lysogeny) before proceeding with infection later. However, many themes are common, including combating host defense systems, shutting off host transcription and translation, appropriating host nutrients and metabolites, and driving necessary metabolic pathways. Next, the viral genome is replicated and packaged into newly synthesized viral particles. Finally, the viral particles extrude from the host cell, or the host cell is physically burst (lysed) open to release as few as two progeny virions or up to tens of thousands. Once outside the host cell, the viral particle will continue to the next host, be actively taken up as nutrients by a non-host, or degrade over time.

Viruses drive evolutionary diversification

Viruses have the unique characteristic of high genomic mutation rates and a propensity for gene transfer[39–42], leading to the ability to evolve at a fast rate. The result is a remarkable pool of genomic and morphologic diversity in the virosphere[1,8,9,43,44]. Besides general adaptation strategies, the fast diversification of viruses aids in overcoming host defense systems or human intervention strategies (e.g., vaccines and anti-viral drugs).

The mutation rate of DNA and RNA viruses has been estimated to be approximately 10^{-8} to 10^{-6} and 10^{-6} to 10^{-4} substitutions per nucleotide per cell infection, respectively[45,46]. Bacteria, in comparison, mutate at a rate of approximately 10^{-10} to 10^{-7} substitutions per nucleotide per generation[47]. With smaller genome sizes and faster generation turnover, viruses can often modify their genomes, in a somewhat randomized fashion, to overcome host defense systems faster than hosts can develop viral resistance. The constant 'battle' between productive viral infection and host resistance has been described as the *viral-host arms race*[48,49]. In addition to promoting expanding viral diversity, it has led to dozens of defense and anti-defense mechanisms, such as CRISPR-Cas, restriction modification, abortive infection, toxin/anti-toxin, and other systems[50–52]. In fact, due to this drive for diversification, viruses have been considered as vital to life processes; viruses are in part necessary for life diversity and evolutionary movement.

Another mechanism by which viruses drive diversification is through horizontal gene transfer. One example is recombination of a host gene with an infecting virus. Here, a virus can acquire a gene from the host and co-opt it for a viral function. Examples of transferring metabolic genes from host to virus are described extensively below. As long as the transferred gene provides a fitness advantage to the virus, overcoming the disadvantage of adding genetic content to a compact genome, the gene can be retained over time[53,54]. Another example is transfer in the opposite direction, where a virus can provide new genetic content to the host. A common

mechanism of this is viral genome integration into the host genome and stable retention of viral genetic content. Some bacteria have obtained virulence factors from viruses, such as *Staphylococcus aureus* and *Vibrio cholerae*[55,56].

Viruses are central to system processes and ecology

Viruses are integral, keystone members of microbial community dynamics, food webs, and nutrient cycles. The main mechanism by which viruses, mostly those infecting microbes, impact global process is by killing their host cell. In fact, it is estimated that 20-40% of all surface ocean microbes are lysed by viruses every day, leading to massive turnover of biomass[37,57]. As a result, viruses have the potential to impact global biogeochemical processes due to the large scale at which viruses influence biogeochemistry of their host and local microbial communities. By killing host cells, viruses contribute in two ways: facilitating diversity and shuttling nutrients.

First, microbial abundances and assortment are controlled. Under a Kill the Winner mechanism, a microbial population that succeeds and grows to high abundance will be effectively sought out by viruses and reduced (killed)[58]. This facilitates dynamic niche differentiation by allowing diverse microbial populations to continuously rise and fall in dominance[59]. Low abundant taxa have periods of proliferation as niches open and nutrient competition is lower. Since microbes are often the main drivers of ecosystem and host (e.g., human) health and nutrient landscape, the control by viruses can have widespread impacts. Microbial viruses have even been identified as markers for human gut health and can be comprised of unique populations between individuals[11,12,60]. In the oceans, viral diversity follows natural gradients, such as depth, oxygen content, specific nutrients, and temperature[10,61].

Second, the cellular contents liberated after host death are available as dissolved and particulate organic matter, in addition to any inorganic nutrients and metabolites. This process enables viruses to be a biomass recycling mechanism[59,62]. In a process termed the 'viral shunt', upwards of 25% of photosynthetically fixed carbon is recycled by viral lysis. Autotrophic (e.g., photosynthetic) microbe, heterotrophic microbe, and higher trophic level grazer biomass is converted into dissolved organic carbon which is in tern taken up by living cells[5]. Moreover, aggregated particles of dead biomass can form sinking particles to recycle nutrients to deeper ocean systems. These repositories of nutrients can also include the viruses themselves, which can account for significant levels of carbon, nitrogen, and phosphorous[63].

Third, viral infection manipulates the internal metabolic landscape of the host cell to drive nutrient acquisition and allocation to virus production[64,65]. The infected host cell (virocell) is necessary to support the carbon, nitrogen, sulfur, and phosphorus needs of the virus. Although viruses can cause the breakdown of host biomass from the inside, such as repurposing host genomic nucleotides for viral genome synthesis[66], the labile nutrients currently within the host are insufficient to meet these demands. To meet the demand, a virocell sustains metabolic activity as seen by the incorporation of medium-derived nitrogen and phosphorus into newly synthesized virions[67–69]. Virocells have also been shown to have influenced and manipulated sulfur and carbon metabolic pathways that are dependent on the infecting virus[70]. Following host lysis, the nitrogen and phosphorus, in the form of excess nutrients or components of viral particles, is released into the environment.

Breaking barriers: the metabolic propensity of viruses

In general, we categorize viruses as non-living due to their inability to self-replicate, lack of ribosomes, and the notion that they are metabolically inert. This barrier is being strained by the identification of viruses encoding ribosomal[71,72] and metabolic genes. Microbes sustain homeostasis and life via complex metabolic networks and nutrient acquisition and turnover. In contrast, viruses may be able to indirectly influence host metabolism as discussed above, but viruses themselves do not necessarily participate in virus driven metabolism. In the early 2000s, evidence to contradict this claim that viruses do not contribute directly to metabolic processes was discovered by identifying photosynthesis genes encoded on viral genomes[73]. It was shown that some viruses infection cyanobacteria can encode core proteins of photosystem II, namely *psbA* and *psbD*, that putatively sustain photosynthetic pathways. Subsequently, it was confirmed that viral *psbA* is actively transcribed during infection, generates protein that integrates into host photosystems, and is integral to viral infection by increasing NADPH (reducing power) for dNTP (nucleotide) production[74–76]. Following this discovery, the capability of viruses to participate in host photosynthesis was identified in a range of systems[77,78].

Photosynthesis and nucleotide metabolism are not the only metabolic pathways that viruses can influence. In 2007, genes encoding metabolic capabilities were given the name auxiliary metabolic genes (AMGs), attributed to the idea that they are auxiliary, but beneficial, to successful viral infection[44]. AMGs can be categorized into two classes: Class I AMGs are directly involved in metabolic pathways (e.g., *psbA*) and Class II AMGs are supportive (e.g., transporters)[36]. Since the first descriptions, there have been AMGs identified for many major metabolisms: carbon fixation (*psbA, psbD*), central carbon metabolism[36], nitrification (*amoC*)[79], nitrogen assimilation[80], methane oxidation (*pmoC*)[81], sulfur oxidation (*dsrA*, *dsrC*)[35,82], phosphorous scavenging (*phoH*)[83], and more[84,85]. It is estimated that viruses acquire AMGs directly from their host during infection,

a process akin to ‘stealing’ metabolic propensity of their living host. Viruses then express and utilize these AMGs to attain a fitness advantage, leading to greater or more efficient viral replication.

The exact fitness advantage depends on the viral replication strategy, host metabolic capabilities, individual AMG, and influenced metabolic pathway. Moreover, viruses often only encode a single AMG, or partial set of AMGs, for any given metabolic pathway. For example, some viruses identified at deep-sea hydrothermal vents have been identified to encode dissimilatory sulfite reductase A (*dsrA*)[82], a major component of dissimilatory sulfate reduction/oxidation. Without context, *dsrA* can contribute to either sulfide oxidation or sulfate reduction, which is dependent on the metabolic capabilities of the host cell. According to sequence homology to *dsrA* of bacteria, since these *dsrA*-encoding viruses likely acquired their AMGs from hosts by lateral gene transfer, the viruses participate in sulfide oxidation. Here, though oxidation of sulfide or stored elemental sulfur, energy (i.e., ATP) is generated and putatively stolen by the virus for more efficient replication. Other key enzymes, such as *dsrB*, have not been identified on any viral genomes. Thus, it is hypothesized that viruses target key, bottleneck components and steps of metabolic pathways for the greatest fitness advantage.

Metagenomics and viromics

Cellular life and viruses exist, in the basic sense, as compartments containing genetic material (i.e., genomes). These genomes contain the instructions for life processes, such as replication and metabolism. Isolating a microbial population and sequencing its genome can yield fundamental information about the way in which the microbes metabolize nutrients, connect with other members of the community, diversify or change in abundance over time, and contribute to

gene transfers. Over time, through biochemical analyses and subsequent genome sequencing, the scientific community has constructed dozens of curated databases of gene annotations to gather information from this genome sequencing.

Early work utilized microbes' 16S small subunit ribosomal RNA sequence (16S) as a marker for taxonomy, abundance, and eventually estimations of metabolism (amplicon sequencing)[86,87]. The 16S gene is favorable because it is well-conserved among members of similar taxonomic groups and is universal in prokaryotes. Using the 16S gene, or even other universal protein-encoding ribosomal genes, microbiomes and systems can be easily and succinctly assessed with relatively low financial or computational burdens. This further led to the construction of large databases of functional annotations for microbial genes, driven by experimental validations, to rapidly analyze the functions of uncultivated microbes in nature.

Rather than the isolation of individual microbes and populations, or sequencing a single gene (e.g., 16S), metagenomics expands on the potential discoveries from genomic information by sequencing whole genomes of an entire community of microbes mixed together. In metagenomics, hundreds to thousands of microbial populations can be sequenced at one time. Metagenomic sequencing can be performed using short-reads, the focus of this dissertation, or long-reads. Extracted genomes are sheared into small fragments and pairs of short reads (e.g., 150bp) are sequenced. Many software tools have thus been developed for analyses of short-read sequencing data. Foremost, assembly can be performed on short reads to re-form respective genomes[88–91]. Often the assembled sequences (contigs or scaffolds) remain highly fragmented, so binning is employed to re-construct metagenome assembled genomes (MAGs) into respective populations[92–97]. From here, there are numerous options for analyzing genomes, such as open reading frame (i.e., gene and protein) prediction[98–100], gene search and functional annotation[101,102], mapping short

reads back onto genomes to obtain relative abundances[103–105], and more. In addition to software development, methods for the physical collection and extraction of genomes from diverse environments have been developed, including filtering microbes from aquatic environments onto a 0.22-micron filter or releasing microbes attached to particles in complex soils.

Following the pioneering of metagenomics for the sequencing of microbial genomes, viromics was established. A virome, analogous to a metagenome, is the collection and sequencing of the viral component of a community or microbiome. Although viruses are often in 10 times greater abundance than microbes in a given system, viral genomes are an average of 100 times smaller than those of bacteria. This leads to disparities in sequencing depth for viruses and microbes, in which more microbial genomes are assembled from a standard metagenome. The disparity is even more evident in filtered samples (e.g., 0.22-micron) because most viruses pass through the filter and are not collected. Viromes allow for the specific sequencing of the viral fraction and can generate distinct, and more robust, information about the viral communities compared to sequencing viruses from a standard metagenome[106].

Computational approaches and methods to study viruses

The expansion of metagenomics and the technology's utility in studying microbial ecology, evolution and metabolism, and subsequent software tool development, has mostly focused on microbes. Viruses pose unique challenges compared to microbes: lack of universally shared genes[107], taxonomic rankings based on morphology rather than sequence similarity[108], less extensive functional annotation databases (i.e., many genes of unknown function), short and fused genes[75,109], highly variable genome sizes (e.g., 4kb to 2500kb)[110,111], the ability to integrate viral genomes into host genomes[112], and more. Future work in viral metagenomics requires methods to

be developed that address these unique challenges and allow for the large-scale study of uncultured viruses. In addition to host prediction and taxonomy, methods that can be improved upon include virus sequence prediction and prophage extraction, prophage activity estimation (i.e., lysogenic to lytic switch), and binning to construct viral MAGs (vMAGs).

Virus sequence prediction is the process of searching for sequences within a metagenome assembly that are of viral origin or for decontaminating a virome assembly. Likewise, prophage extraction is the identification of host genome regions that are integrated viruses (prophages). There are two very general methods for predicting viruses: k-mer patterns and protein annotation. For k-mer patterns, a machine learning model, from random forest decision trees to deep learning long short-term memory networks, is constructed based on observed differences in nucleotide k-mer patterns between viruses and non-viruses[113,114]. For protein annotation, coding sequences are annotated and viral-specific or virus-like patterns are observed[115]. The latter method can also incorporate machine learning[116,117]. Either method employed identifies sequences or sequence regions that are likely to be viral.

Prophage activity estimation is the process of labeling a prophage with its current infection stage, either lytic (active) or lysogenic (dormant). Prophages are unique in that they can exist in a dormant state in which the genome is not be produced for progeny virion formation. However, an integrated prophage continues to be replicated along with the host genome as the host grows and divides. Detecting if a prophage is actively infecting the host and producing progeny genomes versus existing in a dormant state can provide essential ecological information about how a given prophage is impacting the system.

Metagenome or virome assembly often generate fragments of genomes rather than complete genomes. Often, total viral diversity is overestimated due to assuming all viral fragments

represent separate genomes. Constructing vMAGs by binning fragmented genomes allows for more accurate and robust evaluation of viral ecology and metabolism. Yet, binning is not a convention in viral analyses as it is for microbes. Binning vMAGs can be completed in a manner similar to established methods for microbial MAGs, in that sequence feature similarity (e.g., k-mer patterns) and coverage profiles (e.g., relative abundance) between sequence fragments can be used as signatures to re-construct the original genomes[118–120].

Motivation: using bioinformatics to connect viral ecology and metabolism

The evaluation of viruses from metagenome and virome data, such as identifying AMGs or microdiversity patterns, is reliant on a suite of bioinformatics tools. It is essential that such tools are efficient, accurate, and user-friendly for the scientific community. This dissertation follows my research progression through the study of viral ecology and into development of relevant bioinformatics tools and methods. My goal was to apply my foundational knowledge and experiences in microbiology and virology to inform the development of tools that have a strong basis in biological reasoning.

I start by evaluating sulfur metabolism AMGs encoded by viruses, including AMGs of both the assimilatory[121] (Chapter 2) and dissimilatory[122] (Chapter 3) pathways. This includes the global distribution and genomic diversity of the viruses, bioinformatic and experimental evidence of AMG function, and extrapolation to the estimated impacts on microbial community dynamics. These projects highlighted the need for innovations in virus and prophage identification methods, which led to my development of VIBRANT: Virus Identification By iteRative ANnoTation[117] (Chapter 4). Subsequently, I developed PropagAtE: Prophage Activity Estimator[123] to classify prophages as being in the lytic or lysogenic stage of infection (Chapter 5). Finally, at the time of

completing VIBRANT and PropagAtE there was no tool explicitly designed for the binning of vMAGs. Therefore, I created vRhyme, which is a tool and novel method for the binning of viruses from metagenomes or viromes[124] (Chapter 6). Despite the field of viral bioinformatics surging in popularity in the last few years and becoming saturated in bioinformatics tools for some workflows, there is still essential room for improvement. I will conclude this dissertation with my thoughts and opinions on what the future holds for the advancement of bioinformatics methods and conventions for the study of viruses[125] (Chapter 7).

Chapter 2: Virus-associated organosulfur metabolism in human and environmental systems

Kristopher Kieft[1], Adam M. Breister[1], Phil Huss[1,2], Alexandra M. Linz[3], Elizabeth Zanetakos[1], Zhichao Zhou[1], Janina Rahlff[4,5], Sarah P. Esser[4], Alexander J. Probst[4], Srivatsan Raman[1,2], Simon Roux[6], and Karthik Anantharaman[1]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA
[2]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI, USA
[3]Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI, USA
[4]Department of Chemistry, Environmental Microbiology and Biotechnology, University of Duisburg-Essen, Essen, Germany
[5]Centre for Ecology and Evolution in Microbial Model Systems, Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden
[6]Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA

Publication:
Kieft, K. *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep* **36**, 109471 (2021).

All supplementary figures, tables, and files are available at the following Figshare repository:
https://figshare.com/projects/Kristopher_Kieft_PhD_Dissertation/136427

Summary

Viruses influence the fate of nutrients and human health by killing microorganisms and altering metabolic processes. Organosulfur metabolism and biologically-derived hydrogen sulfide play dynamic roles in manifestation of diseases, infrastructure degradation, and essential biological processes. While microbial organosulfur metabolism is well-studied, the role of viruses in organosulfur metabolism is unknown. Here we report the discovery of 39 gene families involved in organosulfur metabolism encoded by 3,749 viruses from diverse ecosystems, including human microbiomes. The viruses infect organisms from all three domains of life. Six gene families encode for enzymes that degrade organosulfur compounds into sulfide, while others manipulate organosulfur compounds and may influence sulfide production. We show that viral metabolic genes encode key enzymatic domains, are translated into protein, are maintained after recombination, and that sulfide provides a fitness advantage to viruses. Our results reveal viruses as drivers of organosulfur metabolism with important implications for human and environmental health.

Introduction

Biological sulfur cycling is one of the oldest and most influential biochemical processes on Earth and is primarily driven by microbial reduction of sulfate to produce hydrogen sulfide [126–128]. Sulfide plays dynamic roles in the degradation of infrastructure and souring of oil reserves [129,130], microbial respiration and essential biosynthesis processes, and manifestation of human gastrointestinal disorders such as colitis, inflammatory bowel diseases (IBD) and colorectal cancer (CRC) [131]. Much of our knowledge of sulfur cycling focuses on a small subset of microbes that are capable of respiring and transforming inorganic sulfur compounds, a process known as

dissimilatory metabolism [132]. Consequently, the cycling of sulfur-containing organic (organosulfur) compounds and resulting sulfide production from more widespread biological mechanisms and sources has largely been ignored.

Two mechanisms of sulfide production include the degradation of organosulfur compounds and assimilatory sulfur metabolism. Sulfide production from microbial-driven degradation of organosulfur compounds, such as the amino acid cysteine, has been noted as a significant contributor to sulfide concentrations in environmental and human systems [133–135]. However, there exists no comprehensive analysis of the specific microbes involved. Assimilatory sulfur metabolism, a common strategy used by many microbes and some eukaryotes to incorporate sulfide into biological compounds, has similarly been routinely discounted as a mechanism of significant sulfide release into either environmental or human systems. Notably, the role of viruses in these processes has not been explored.

Microbial viruses, mainly comprising bacteriophages (phages) are extraordinarily abundant on Earth. Microbial viruses are known to redirect and recycle nutrients on the scale of ecosystems by infecting and lysing host cells [5,63,136,137]. In the oceans alone, the number of viral infections per second exceeds the number of stars in the known universe, which likely leads to the lysis of over 20% of all microbes per day [37,138]. In addition to lysis, viruses can actively redirect host metabolism during infection which manipulates major biogeochemical cycles, including carbon, nitrogen, and sulfur. One such mechanism involves viruses “stealing” metabolic genes from their host in order to gain fitness advantages during infection [78]. Such host-derived viral genes are termed auxiliary metabolic genes (AMGs), and are expressed during infection to modulate microbial respiration, biosynthesis processes, and/or direct intracellular nutrients towards virus replication and virion production [6,38,44,73,75,82,84,122,139,140]. For example, some viruses of

Cyanobacteria encode core photosystem proteins that augment host metabolism in order to increase the biosynthesis of dNTPs that are utilized for viral genome replication [75]. The viral auxiliary metabolism of iron-sulfur clusters, central carbon metabolism, nitrification, methane oxidation, and other metabolic processes could also provide viruses with a multi-faceted method of manipulating nutrients within their host cell to enable efficient, rapid or otherwise a more improved viral replication cycle [36,79,81,139].

In spite of the importance and global prevalence of viruses, nothing is known about their contribution and impact on AMG-driven organosulfur metabolism in the environment. Moreover, the role of AMGs in human microbiomes has been largely unexplored. Here, we investigated environmental and human microbiomes for the presence of viruses involved in production of hydrogen sulfide and manipulation of organosulfur metabolism. By screening publicly available partial and complete viral genomes from cultivated and uncultivated viruses, we identified genes involved in direct and indirect sulfide production from organosulfur degradation and assimilatory sulfur metabolism. We followed this up with experiments to validate the impacts of genes for organosulfur metabolism as well as hydrogen sulfide on viral fitness.

Results

Metabolic pathways for organosulfur metabolism driven by viral AMGs

We queried a comprehensive dataset of approximately 135,000 partial and complete viral genomes (contigs) publicly available on Integrated Microbial Genomes/Viruses (IMG/VR) [9,141] and the National Center for Biotechnology Information (NCBI) databases, and two metagenomic studies from Lake Mendota, WI [142], for the presence of virally encoded proteins for organosulfur metabolism. In total, we identified 4,103 viral AMGs representative of 39 unique gene families.

All genes identified are categorized as Class I AMGs, or those for central metabolic functions but auxiliary to productive viral infection [36]. These AMGs were detected on 3,749 non-redundant viral genomes from all major bacterial dsDNA viral families (*Myoviridae, Podoviridae* and *Siphoviridae*) including viruses infecting an archaea [143] and eukaryote (amoeba) [144]. Therefore, AMGs for organosulfur metabolism were identified on viruses infecting all three domains of life, representing a shared metabolic constraint regardless of host domain. The viruses represent cultivated and uncultivated viruses, linear and circular genomes, and lytic and lysogenic cycles of viral replication across a vast range of environmental and human microbiomes. Of these, 164 have been isolated and cultivated on hosts spanning nine major bacterial lineages (Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Cyanobacteria, Actinobacteria, Firmicutes, Bacteroidetes, Verrucomicrobia and Deinococcus-Thermus) as well as an amoeba (*Vermamoeba vermiformis*) (**Table S1**). The isolation of viruses encoding organosulfur metabolism AMGs

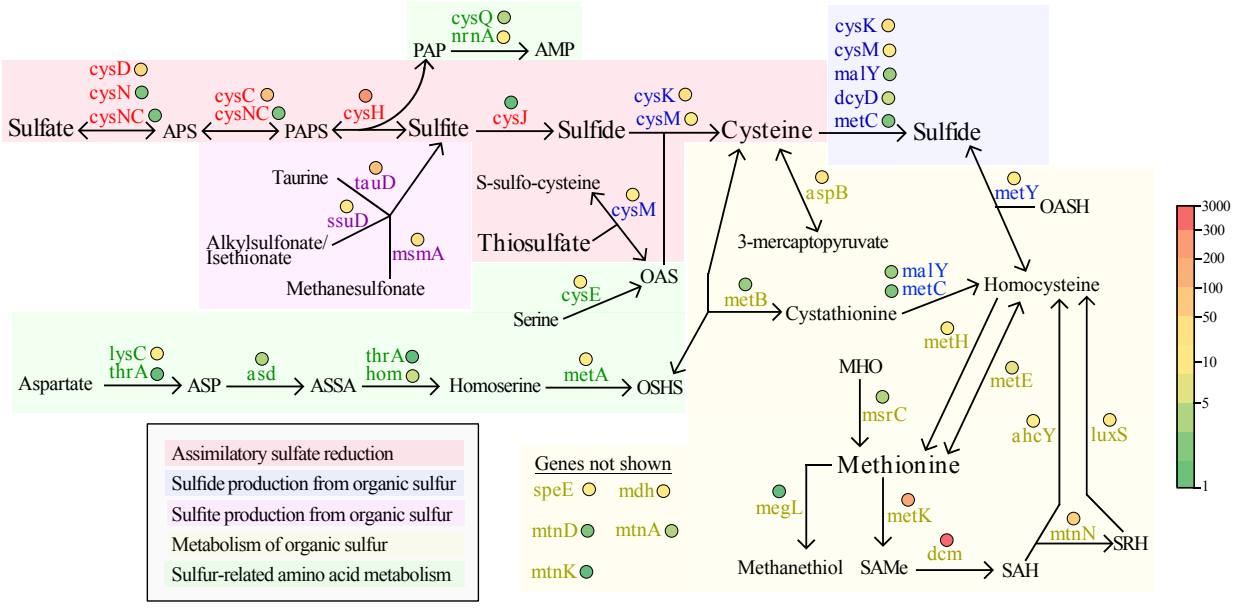


Figure 1. Reaction diagram of organosulfur transformations mediated by viruses. All genes shown have been identified on viruses and are colored coordinated respective to the process with which they are putatively associated. Colored circles represent the abundance of each AMG across all viral genomes according to the color scale (heatmap) on the right. Complete reactions and full names of acronyms are listed in Table 1.

indicates that the identification of such viral driven metabolism is not an artifact of metagenomic analysis.

Viral AMGs are putatively associated with five distinct processes: sulfide production from organic sulfur, the assimilatory sulfate reduction pathway, sulfite production from organic sulfur, metabolism of organic sulfur, and sulfur-related amino acid metabolism (**Figure 1 and Table 1**). Six different AMG families (*cysK*, *cysM*, *malY*, *dcyD*, *metC* and *metY*) encode for enzymes able to directly produce sulfide from the degradation of cysteine and homocysteine, which are important organosulfur compounds and central sources of sulfur in the environment and human body [145,146]. Six other AMG families (*cysD*, *cysN*, *cysC*, bifunctional-*cysNC*, *cysH* and *cysJ*) are components of the assimilatory sulfate reduction pathway, which is widely utilized across all three domains of life for incorporation of sulfide into cysteine. Sulfite can be directly produced from the breakdown of several organosulfur compounds (e.g., taurine) by three families of AMGs (*tauD*, *ssuD* and *msmA*) and successively fed into dissimilatory and assimilatory sulfate reduction. Eleven of the AMG families (*aspB*, *metB*, *metH*, *metE*, *msrC*, *metK*, *megL*, *dcm*, *mtnN*, *ahcY* and *luxS*) are inferred to indirectly produce sulfide by manipulating abundant organosulfur compounds (e.g., methionine and cystathionine) that funnel into the synthesis of cysteine or homocysteine. Finally, indirect organosulfur metabolism by the remaining thirteen AMG families (*lysC*, *thrA*, *asd*, *hom*, *metA*, *cysE*, *cysQ*, *nrnA, speE, mdh, mtnD, mtnA* and *mtnK*) would influence the synthesis of organosulfur compounds (e.g., synthesis of cysteine using serine) that feed into sulfide producing reactions.

Viruses encoding AMGs for organosulfur metabolism are globally distributed

Uncultivated viruses encoding AMGs for organosulfur metabolism were recovered from diverse environmental (marine, freshwater, engineered, soil, hydrothermal vent, non-marine saline and alkaline, deep subsurface, wetland and thermal spring), non-human host-associated (mammalian gut, other animal-associated and plant-associated) and human host-associated (gastrointestinal, oral and vaginal) microbiomes (**Figure 2A**). Cultivated and well-characterized viruses exhibited likewise microbiome dispersal because they were recovered from more than one ecosystem (e.g., food production, marine, freshwater, soil, engineered, hot springs, animal-associated, plant-associated, as well as human-associated gastrointestinal, oral and skin) (**Table S1**). These results encompassed every ecosystem category, with the exception of air, in which viruses are routinely identified. This displays evidence that viruses encoding AMGs for sulfide production are ubiquitous on Earth.

Next, we estimated the proportion of viral richness in each ecosystem category found to encode organosulfur metabolism AMGs. Viruses encoding at least one AMG were found to be highly abundant in human vaginal, gastrointestinal and oral microbiomes comprising 8%, 6% and 3% of all identified viruses, respectively. Mammalian-associated, other animal-associated and plant-associated microbiomes likewise had significant AMG-encoding virus abundances of 8%, 6% and 6%, respectively. Notably, previous reports have determined that expanded viral richness in the gastrointestinal tract is correlated with the manifestation of IBD [147] and our results support the possibility of this being in part due to the metabolic potential of viruses, such as for sulfide production. This points to an important distinction that the collective metabolic potential of viruses in these host-associated environments, in conjuction with measuring total viral richness, could have significant implications for host health. Viruses encoding organosulfur AMGs beyond host-associated microbiomes may also impact ecosystem health. Major environmental systems, such as

the deep subsurface (6%), engineered (3%), soil (3%), freshwater (2%), wetlands (2%), marine (2%) and hydrothermal vents (2%), likewise display significant richness of organosulfur AMG encoding viruses (**Table S2C**). The net impact of viral metabolism on organic and inorganic sulfur

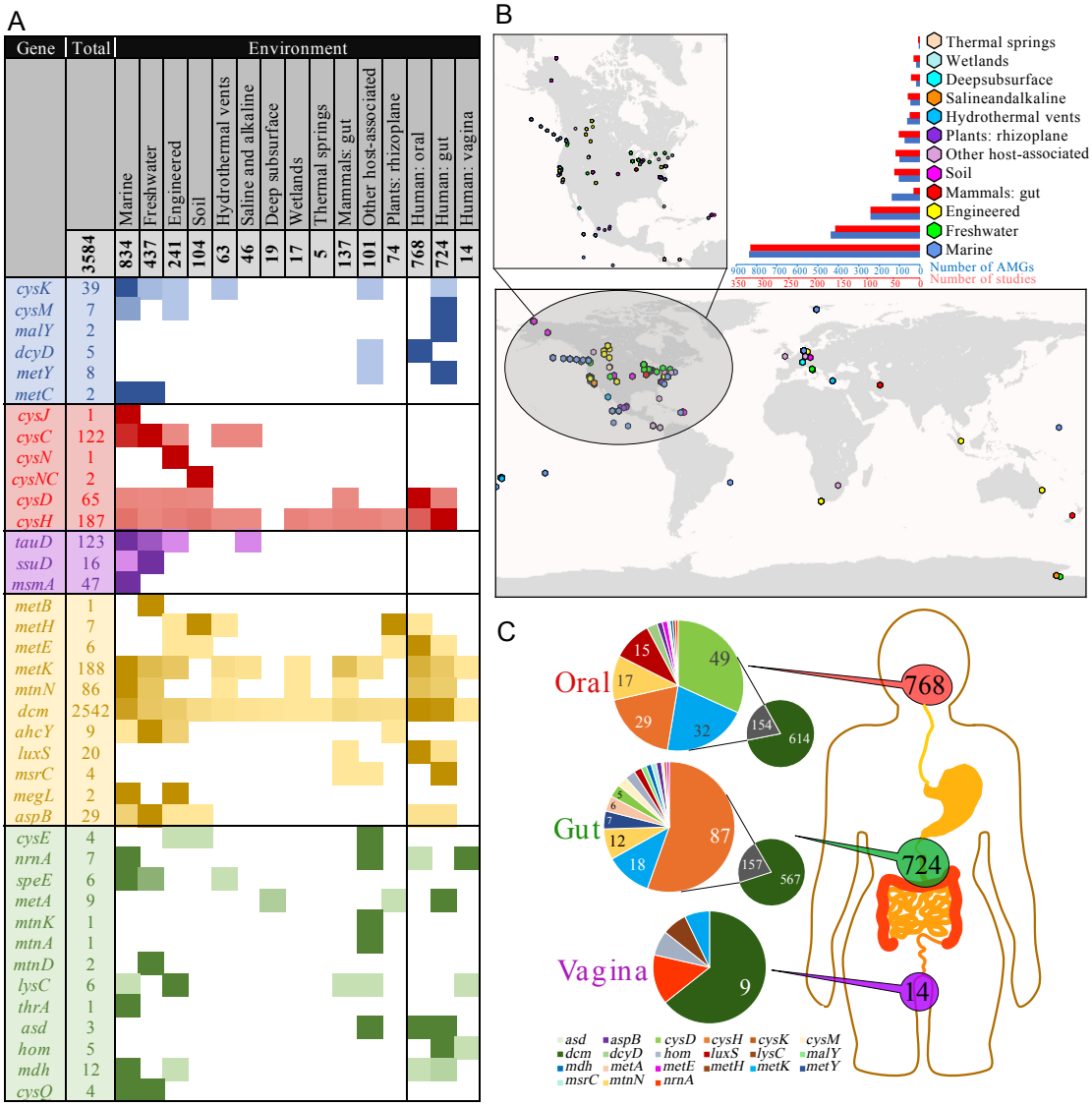


Figure 2. Distribution of viral AMGs in environmental and human microbiomes. (A) Heatmap of each AMG's relative abundance in environmental and human systems with colors coordinated by the AMG's pathway respective to Figure 1. Per AMG, darker colors represent greater abundance. A total of 3,584 AMGs derived from IMG/VR are shown. (B) Global distribution of viruses encoding AMGs, color coordinated by environment classification. The bar graphs represent the number of AMGs and IMG studies from which viruses were recovered. See Tables S2A and S2B for exact abundances for (A) and (B), respectively. Only studies with published coordinates and environment categories are shown. (C) Abundance of AMGs derived from incomplete or uncultivated viruses from human oral, gastrointestinal and vaginal microbiomes. Only values greater than five are shown.

compound concentrations in these environments is unknown, but it is nonetheless striking that up to 8% of all resident viruses may be involved.

Viruses recovered from non-human microbiomes also displayed extensive geographical and niche distributions, which demonstrates their relevance in global sulfur biochemistry (**Figure 2B**). Individual distributions of abundant AMGs (e.g., *dcm*, *cysC*, *cysK*, *cysH*, *metK,* and *tauD*) likewise had no geographical or environmental restrictions (**Figure S1A-F**). For example, *cysH* which encodes a critical enzyme for assimilatory sulfur metabolism was found in every ecosystem except the deep subsurface. CysK, a predominant enzyme involved in sulfide generation from cysteine degradation was also broadly dispersed in marine, freshwater, engineered, hydrothermal vent and host-associated environments. Even *msmA* which was only identified in marine environments showed strong geographical dispersal (**Figure S1G**).

AMG distributions between environments may depend on different factors, such as how universal the AMG function is (e.g., CysH and CysK are common amongst bacteria) or the nutrient landscape in a specific environment (e.g., MsmA is capable of degrading methanesulfonate, a common compound in marine environments [148]. However, human-associated samples contained the greatest fraction of identified *cysH* and *cysD* AMGs overall, while marine and freshwater environments contained nearly all of the identified *cysC*. In human-associated samples, nearly 97% of AMGs were *cysD*, *cysH*, *metK*, *mtnN*, *luxS* and *dcm* which encompass essential steps of cysteine and methionine degradation (**Figure 2C**). The uneven distribution of these assimilatory sulfate reduction AMGs suggests that further constraints on nutrient availability or variance in rate limiting steps based on thermodynamics in different environments play a role in determining the distribution of organosulfur metabolism AMGs.

Viral organosulfur AMGs result in likely functional proteins and provide a fitness advantage to the virus

To overcome the challenge of assigning conclusive function to protein sequences in the absence of biochemical evidence, we analyzed functional and conserved domains of AMG-encoded proteins with biochemically characterized bacterial homologs. Overall, we examined 24 AMG families and found broad conservation of whole protein sequence and functional amino acid residues (**Figure S2**). For example, viral sequences encode specific domains for: CysC: ATP binding (gsGKss) and required motifs (dgD) [149]; CysK: cofactor pyridoxal phosphate binding (KDR, NtG, GT/SgGT and SS/AG), substrate binding (T/SSGN and QF) and phosphate recognition (GI/V) [150]; MetK: substrate binding (egHPDk, acE, gEit, GDqG, DaK, TgRKi, sGKd and kvDrs) [151]; CysH: iron-sulfur cluster motif (CC…CxxC) [152]; TauD: nitrogen and oxygen binding (e.g., WH and H) [153]. Conserved amino acid residues that are not functional are likely preserved for structural features. The retention of AMGs on viral genomes despite strong selective pressures for reduced genome size suggests that most of these AMGs are functional [74]. In addition to functional and conserved domain analysis we calculated the ratio of non-synonymous to synonymous nucleotide differences (*dN/dS*) for a subset of the abundant viral AMG families. A *dN/dS* value less than one would suggest that the virus is under selective pressures to retain a functional AMG. *dN/dS* calculations for *cysK, cysC, cysD, cysH, tauD, msmA, metK, mtmN and luxS* AMG pairs revealed that viral AMGs appear to be under purifying selective pressures to retain function of the encoded AMGs (**Figure S3**).

To assess if viral AMGs are active in the environment, we queried a comprehensive metagenomic and metatranscriptomic dataset from Lake Mendota, WI. We identified 23 AMGs representative of six gene families (*aspB*, *cysC*, *cysH*, *metK*, *speE* and *tauD*) that were actively

expressed by 22 different viruses over a 48-hour time period (**Table S3**). One *cysC* in particular was expressed by a virus with a 210kb genome that was bioinformatically determined to be complete and circular. Analysis of the genome's GC-skew, a metric to evaluate genome replication patterns using nucleotide coverage [154], was used to determine that the virus performs rolling circle replication (i.e., unidirectional) which is a common method utilized by viruses [155] (**Figure S4A**). To assess if the virus was actively replicating when *cysC* was expressed we used a metagenomic read mapping approach to estimate the genome's *in situ* index of replication (iRep) [156]. The genome's iRep value of 1.54 falls within the range of typical values of growing populations and indicates that the virus was actively replicating its genome in the environment when *cysC* was expressed (**Figure S4B**). Analyses of other host-virus systems with transcriptomic data enabled the identification of *cysH* expression by Enterobacteria phage Lambda during infection of *Escherichia coli* MG1655 [157]. The activity and expression of viral AMGs in various systems provides further evidence that they are likely utilized for a specific function during infection.

To validate that AMGs are in fact transcribed during infection we developed a model host-virus system with *Lactococcus lactis* C10 and its *cysK*-encoding virus Lactococcus phage P087. The transcript abundance of *cysK* was measured in a culture of either *L. lactis* C10 grown alone (control) or with P087 at timepoints 15-, 60- and 120-minutes post infection (**Figure S5** and **Table S4A**). At 120 minutes the host cells in the infection condition had mostly lysed from viral infection. Transcript abundance of *L. lactis* C10 *cysK* was found to be comparable at 15 minutes and 60 minutes in either the uninfected control or infected with P087. At 120 minutes transcripts of *L. lactis* C10 *cysK* were 4x greater than at 60 minutes in the control but were undetectable in the infected condition. This suggests that *L. lactis* C10 *cysK* transcripts are greatly reduced during mid to late infection by P087. The transcript abundance of P087 *cysK* follows a similar trend as *L.*

lactis C10 *cysK*. At 15 minutes P087 *cysK* transcripts were near zero and by 60 minutes were in approximately 2x greater abundance compared to transcripts of the host. By 120 minutes P087 *cysK* transcripts likewise reduced nearly to initial levels. There was no detection of P087 *cysK* transcripts within the uninfected control. Although P087 *cysK* transcript abundance never exceeded that of *L. lactis* C10 *cysK*, we provide further evidence that the viral AMG *cysK* is actively transcribed during infection and potentially replaced host *cysK* to an extent with the greatest abundance during mid infection rather than early or late infection.

To validate that transcribed AMGs in fact produce protein, we further leveraged the *L. lactis* and P087 system. Using untargeted mass spectrometry at the endpoint of virus infection (i.e., lysis) we identified that P087's AMG *cysK* produces protein and at approximately 1.5x greater abundance than *L. lactis* C10 *cysK* (**Table S4B**). The higher ratio of virus CysK to host CysK suggests the virus gains a fitness advantage from compensation of CysK levels in the cell. These findings build upon the results from our qPCR-based analysis of transcript abundance in which host transcripts were more abundant than viral but may be explained by higher stability of either viral CysK or *cysK* transcripts. Moreover, since viruses demand a substantial fraction of cellular resources during infection [53], the high viral CysK levels measured here supports our hypothesis that CysK is actively utilized during productive infection in contrast to being metabolically inactive. The presence of the gene on the genome in conjunction with transcription and translation measurements is consistent with the AMG providing a fitness advantage, which has been modeled to be as much as a 4% gain for some AMGs [74]. The mechanism(s) by which this functions is likely different than what has been observed previously for AMGs. For example, AMGs for photosynthesis were found to have differential effects during light-dark cycles as well as transcript compensatory effects over an ~8 hour time period [75]. Conversely, P087 is not influenced by light-

dark cycles and complete lysis can occur within ~2.5 hours. Beyond providing evidence that AMGs can be remarkably active during infection this further underlines the diverse nature by which AMGs are utilized by viruses. In addition, the identification of similar gene families on genomes of diverse, geographically spread viruses strongly supports the hypothesis that organosulfur metabolism AMGs play a functional role during infection [140].

Viruses encoding organosulfur AMGs are phylogenetically diverse

To investigate the diversity of AMGs we conducted phylogenetic analysis of encoded amino acid sequences for five gene families. Phylogeny of CysH from complete viral genomes show close relationships between viruses and their known hosts, supporting previous observations that AMGs are most often acquired from the host [78] (**Figure S6A**). One clade in particular encoded an addition domain of unknown function (DUF3440) which suggests a shared evolutionary history. Analysis of CysH phylogeny of viral contigs with no known host revealed a similar clustering of viruses with their putative bacterial hosts (phyla Bacteroidetes and Firmicutes) (**Figure S6B**). In contrast to CysH, phylogenetic analysis for several abundant AMG protein sequences (CysC, CysK, TauD and MetK) on complete and incomplete viral genomes displayed clustering of viral sequences in separate clades from bacterial homologs with few exceptions of the virus clustering with a putative host (**Figure S6C-F**).

Separate clustering would suggest that viruses may have acquired AMGs beyond their current or known host range, which is supported by the observation that viruses can encode an AMG that their host does not (e.g., *cysC* for *Xylella* phage Sano) and that AMGs can cluster separately from their host (e.g., CysH for *Vibrio* phages). However, based on the CysH phylogeny of complete viral genomes another likely explanation for distinct viral clustering is that the full

range of host sequences has yet to be identified. Within the human microbiome alone, thousands of novel bacterial genomes have been identified recently and may provide further insight into host ranges or origins of AMG transfer [158–160]. Even so, in comparison to human microbiomes, little is known about the breadth and diversity of environmental or human viromes. Analysis of all AMGs suggests they have collectively been derived from bacteria (with the exceptions of the archaeal and eukaryotic viruses) affiliated with the phyla Firmicutes, Bacteroidetes, Alphaproteobacteria and Gammaproteobacteria, which is supported by the host range of cultivated AMG-encoding viruses (**Table S1**).

Directed recombination and AMG sequence conservation validates proposed mechanism of AMG transfer and retention

The proposed mechanism of AMG acquisition by viruses in nature is the transfer of a host metabolic gene to the virus by recombination. Over multiple replication cycles of the viral genome, the AMG is retained as a functional gene. To verify this proposed mechanism, we engineered *Escherichia coli* phage T7 by inserting the host gene *cysK* (T7::*cysK*) to simulate a recombination event. Following successful insertion, T7::*cysK* was passaged, in three biological replicates, for nine complete infection cycles to simulate infection in nature over time. After passaging, the T7::*cysK* construct was sequenced to check for retention of the AMG in the viral population. Sequencing confirmed retention of the gene, indicating that recombination of a host metabolic gene onto a viral genome (i.e., AMG acquisition) can lead to stable retention of an AMG over time. Furthermore, between three biological replicates no mutations from the wildtype *cysK* sequence were observed.

Importantly, these observations show that a recombination event can occur without environmental triggers (e.g., nutrient limitation during infection) or fitness constraints (e.g., metabolic bottlenecks in the host), which provides further credibility for the proposed mechanism that AMG transfer occurs frequently and randomly in nature. If the AMG provides sufficient fitness benefits, or a lack of detrimental effects on viral replication it will be retained over multiple infection cycles. In the system developed here, conditions resulting in a fitness benefit (e.g., greater burst size or faster replication) for the T7::*cysK* virus compared to wild-type T7 were not identified.

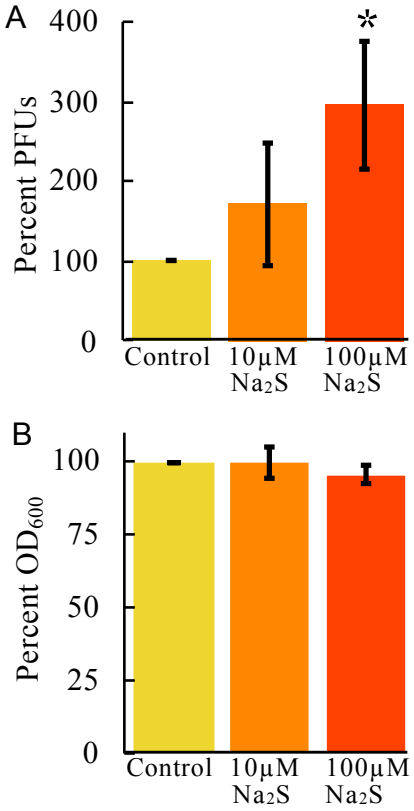


Figure 3. Increased viral fitness is associated with sulfide concentrations. Impact of varying sulfide concentrations on (A) Lactococcus phage P087 virus production as measured by plaque forming units (PFUs). The average of three independent experiments each with three biological replicates, with standard deviation error bars, are shown. (B) Corresponding uninfected host growth as an average of two biological replicates. Experimental conditions are normalized to percent of control. Asterisk represents statistical significance ($p<0.02$) compared to the control.

Sulfide can provide a fitness advantage to viruses

Since active expression and function of AMGs likely can result in the production of sulfide in the environment and human microbiome, we sought to determine if sulfide does indeed confer a fitness advantage to viruses. A highly plausible method for viruses to achieve this would be through the degradation of cysteine which is present in nearly all environments. As a result, we hypothesized the *cysK*-encoding virus P087 would have the capacity to gain a fitness advantage in the presence of sulfide. Theoretically P087 would be involved in the direct

degradation of intracellular cysteine via the action of virally encoded CysK under some conditions. To elucidate if sulfide alone confers a fitness advantage, we exogenously added sulfide during P087 infection of *L. lactis* and quantified the impact on virus and host growth. We found that viable virus production increased linearly with the addition of physiologically relevant concentrations of sulfide (**Figure 3A**) with no significant observed differences in host growth (**Figure 3B**). This indicates that under the conditions tested P087 benefits from increased production of sulfide in the system through either AMG or host-driven mechanisms, and that the resulting fitness gain is not due to a simple increase in host abundance. We performed the same experiment with exogenously added cysteine but did not observe any effect on viral fitness (data not shown). This has significant biological implications as microorganisms contain high intracellular concentrations of cysteine, with *L. lactis* species reported to contain approximately 3.5mM intracellular cysteine [161]. Likewise, *Escherichia coli* has a free cysteine pool of approximately 150μM [162]. We believe other viruses encoding organosulfur metabolism AMGs would likewise derive a fitness advantage under similar conditions and that this phenotype is not restricted to the ability to directly produce sulfide from cysteine degradation.

Viral organosulfur auxiliary metabolism associated with human gut bacteria

Among viruses with known hosts, 107 were found to be associated with 35 different bacterial species known to be commensal or pathogenic residents of the human gastrointestinal tract (**Table S1**). These viruses encode five AMGs (*cysE*, *cysH*, *cysK*, *dcm* and *metK*) for both the assimilation of sulfur and capacity to degrade organosulfur compounds into sulfide. Most of these viruses were isolated from a variety of dairy, soil, sewage, and wastewater environments indicating a potential for environmental reservoirs of sulfide producing viruses, or in the case of wastewater

environments the viruses may have been resident in human gastrointestinal tracts. Five AMG-encoding viruses of the pathogens *Salmonella enterica*, *Staphylococcus aureus*, *Vibrio cholerae* and *Clostridium difficile* were isolated from human fecal samples indicating transmission and replication in human gastrointestinal tracts likely does occur and may contribute to dysbiosis via the production of sulfide or altering the organosulfur metabolic potential of the pathogenic host.

Uncultivated viruses from the human gastrointestinal tract encoding AMGs putatively involved in direct sulfide production (*cysM*, *malY* and *metY*) had high protein identity (>97%) to *Alistipes putredinis, Alistipes obesi, Alistipes finegoldii, Bacteroides uniformis* and *Bacteroides vulgatus* suggesting they are viruses closely associated with these human gut bacteria from the order *Bacteroidales* (phylum Bacteroidetes) [163–166]. Viruses encoding *metK*, *mtnN* and *metE* (i.e., capacity for methionine degradation to sulfide) in human gastrointestinal samples were likewise inferred to be closely associated with the human gut bacteria *Alistipes ihumii*, *Faecalibacterium prausnitzii, Flavonifractor sp., Bacteroides intestinalis, Bacteroides xylanisolvens, Bacteroides uniformis, Bacteroides thetaiotaomicron, Haemophilus parainfluenzae, Aggregatibacter sp*. and *Eubacterium sp*. based on high protein identity [167–175]. At lower protein identity (96%-80%), viruses encoding *metK*, *luxS* and *mtnN* were inferred to be in some part associated with the gut bacteria *Prevotella spp.* (*Bacteroidales*), *Butyricicoccus spp.* and *Clostridiales sp.* [165,176,177] (**Table S5**).

Many of these *Bacteroidales* (i.e., *Alistipes spp.*, *Bacteroides spp.* and *Prevotella spp.*) and some members of the phylum Firmicutes (e.g., *Haemophilus parainfluenzae* and *Butyricicoccus spp.*) have been strongly associated with IBD [166,175,176,178] and their role in inflammation may be in part attributed to virus-mediated or influenced production of sulfide. Importantly, viruses of these *Bacteroidales*, including *Prevotella* megaphages with high coding capacity, have been shown to

be dominant and abundant in human gastrointestinal tracts which could promote the continuous viral-driven production of sulfide to exacerbate inflammation [179,180].

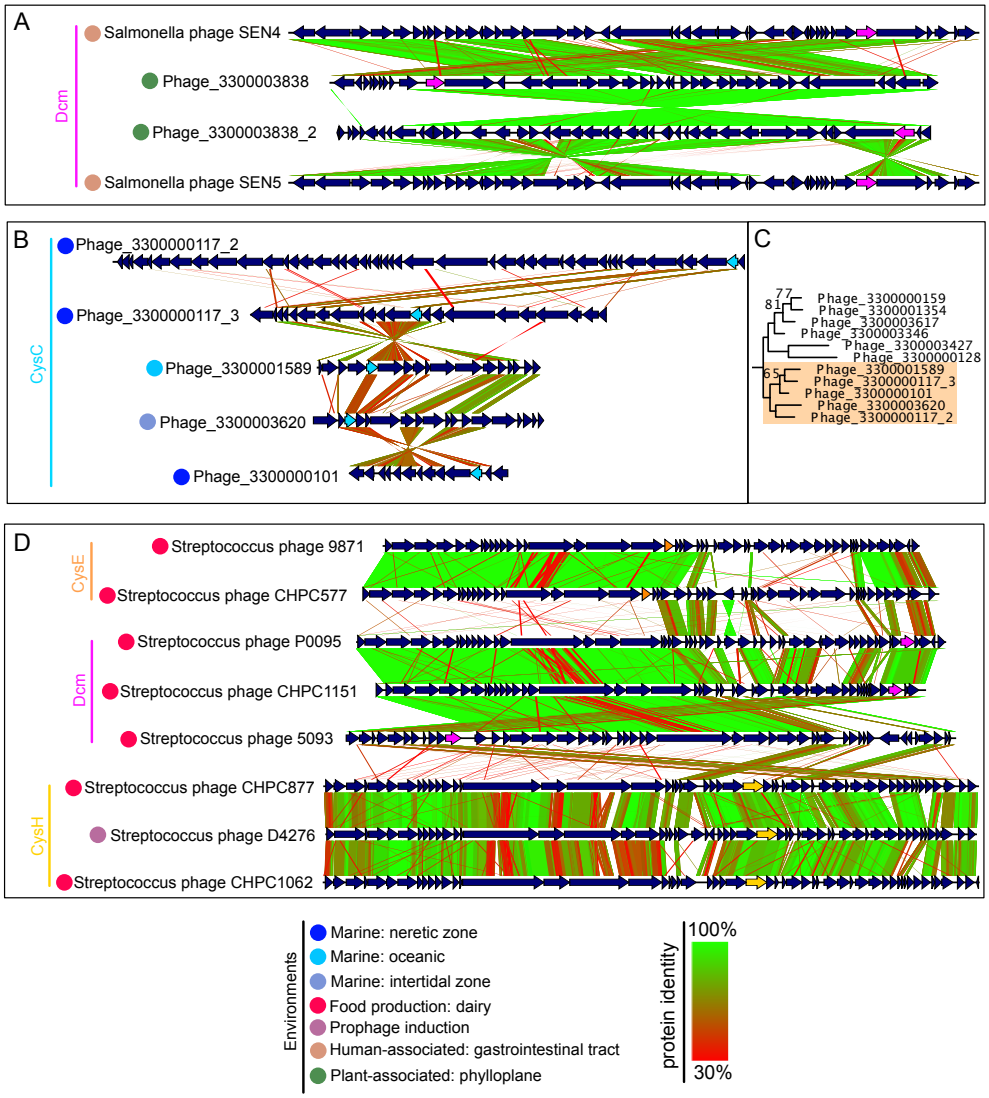


Figure 4. Genome comparisons of viruses encoding AMGs. Comparisons of (A) uncultivated viruses and complete *Salmonella enterica* viruses encoding *dcm* (pink), (B) uncultivated viruses encoding *cysC* (cyan) with (C) respective protein phylogeny (orange highlighting, refer to Figure S6 for full phylogenetic tree), and (D) complete *Streptococcus thermophilus* viruses encoding *cysE* (orange), *dcm* (pink) or *cysH* (yellow). For all comparisons, predicted open readings frames are annotated by dark blue arrows and genomes are connected with lines according to protein identity by tblastx alignment. Colored circles refer to the environment in which the virus was isolated or identified.

Comparative genomics displays diversity of viral genome organization

We used comparative genomics to examine the diversity of viruses found to be associated with human microbiomes. We identified four distinct uncultivated virus contigs encoding *dcm* from human oral samples to be closely related to known *Streptococcus pneumoniae* viruses based on genome sequence identity (**Figure S7A**). However, there are large stretches of dissimilarity between some of the genomes which may indicate evidence for large genetic exchange between viruses that frequently share the same niche and not the same host, which has been demonstrated before between *Lactococcus* and *Enterococcus* viruses [181]. This observation supports the likelihood of AMG transfer between viruses in human and environmental microbiomes. Furthermore, two plant-associated viruses were identified to be closely related to known *Salmonella enterica* viruses originally derived from human fecal samples (**Figure 4A**). These plant-associated viruses may represent examples of environmental reservoirs for AMG-encoding viruses in the human gastrointestinal tract.

However, for either case above the exact nature of viral transfer of AMGs is challenging to determine because AMG sequences that closely share evolutionary history can be encoded on dissimilar and geographically diverse viruses. For example, five *cysC*-encoding viruses that group closely by CysC phylogeny conversely depict dissimilarity by genome comparison and are geographically dispersed in marine environments (**Figure 4B, C**). The same is true for six different *metK*-encoding viruses in which MetK shows phylogenetic similarity but the genomes are diverse and geographically spread (**Figure S7B**).

To further investigate the relationships of AMGs on viral genomes, we examined the prevalence of multiple AMG copies on individual genomes. In total we identified 285 viral genomes that contained multiple copies. While most such genes encoded for identical functions

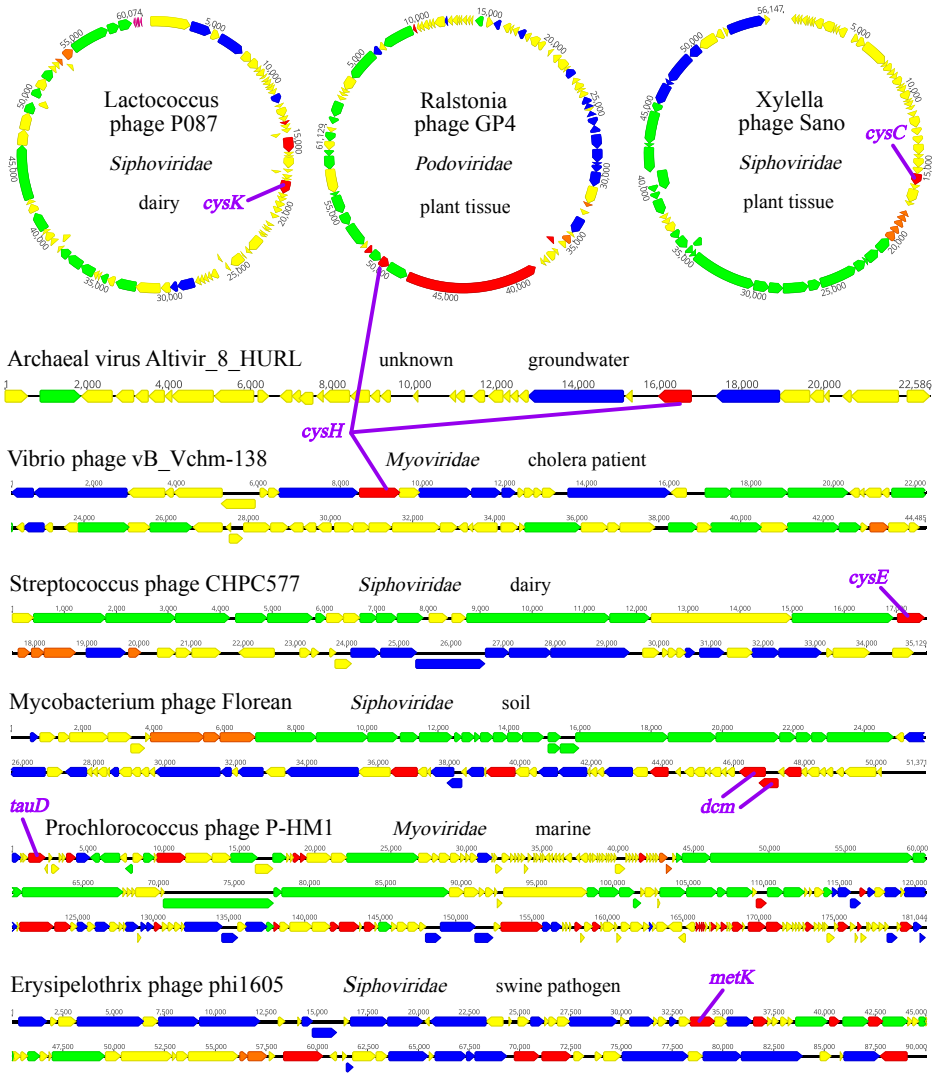


Figure 5. Genome organization of 9 complete viral genomes encoding organosulfur AMGs. Genome representation of circular and linear viruses. Arrows indicate open reading frames and are annotated by general function: virion structural assembly (green), auxiliary metabolism and general functions (red), nucleotide metabolism and genome replication (blue), lysis (orange) and unknown function (yellow). AMGs are annotated in purple.

(i.e., two copies of protein from the same gene family), some with connected (e.g., *metK* and *dcm*, *luxS* and *mtnN*) or disparate functions (e.g., *dcm* and *cysM*, *cysH* and *mtnN*) were also found. These findings suggest viruses may utilize these genes for diverse regulation of host organosulfur metabolism to fit their individual requirements (**Table S6**). For example, a single virus may augment both assimilatory sulfate reduction (e.g., using CysH) as well as methionine degradation (e.g., using MetK) during infection by encoding and expressing both AMGs.

We next compared viral genome organization to identify relationships in the physical location of AMGs between different viral genomes and interpret affiliations with other encoded genes. We found no universal organization of AMGs which were broadly encoded in various locations, such as between structural genes, adjacent to lysis factors, near genes for genome replication or nucleotide metabolism and within regions comprising genes of unknown function (**Figure 5**). Additionally, no pattern associated with encoding specific AMGs was detected according to virus classification, genome length or isolation source. There were a small number of outliers, such as a comparison of 10 complete viral genomes encoding *cysH* that indicated a trend towards co-location of the AMG with genome replication and/or nucleotide metabolism genes to suggest similar transcriptional regulation or function of this AMG across different viruses (**Figure S7C**).

The model that viruses acquire AMGs from diverse sources and for disparate functions is further supported by looking at AMG-encoding viruses that share the same host but not the same AMG. There are several different variations in which this occurs. One example involves *Bacillus cereus* phages PBC5, Basilisk, BCU4 and PBC6 where the viruses have low sequence similarity between genomes and AMG sequences (i.e., *cysH*) (**Figure S7D**). Another example involves *Streptococcus suis* phages phiJH1301-2, phiSC070807, phiNJ3 and phiD12 where the viruses have very similar genome sequences but encode multiple AMGs with similarity shared only among a subset of them (i.e., *metK* and *dcm*) (**Figure S7E**). A final example involves *Streptococcus thermophilus* phages 9871, CHPC577, P0095, CHPC1151, 5093, CHPC877, D4276 and CHPC1062 where the viruses group separately according to the single AMG each encodes (*cysE*, *cysH* or *dcm*) (**Figure 4D**). Taken together, these three examples indicate that viruses are able to employ separate strategies to accomplish a similar function of manipulating host organosulfur

metabolism. This may be in the form of acquiring the same AMG from different sources to perform a shared task or acquiring disparate AMGs to perform separate tasks towards the same objective, such as sulfide production.

Discussion

The metabolic potential of viruses, the most abundant biological entities on Earth, is all too often overlooked because viruses do not independently conduct metabolic transformations. Here we show that viral manipulation of host metabolism in contrast to solely measurements of viral richness and host range is likely important to the environmental sulfur cycle and human health. Furthermore, we propose that assimilatory sulfur metabolism, a ubiquitous method of fixing sulfur and manipulating organosulfur compounds, is frequently modulated by viruses during infection of organisms from all three domains, and in almost all microbiomes on Earth. This poses an important question, what have we been overlooking in viromes by frequently assessing sequence reads instead of metagenomically assembled genomes that encode AMGs? Are we giving enough emphasis on viruses as core drivers in the metabolism of microbiomes?

AMG-driven organosulfur metabolism mediated by viruses may lead to sulfide production in the gastrointestinal tract during infection or following microbial lysis. The result would be a sulfide-induced inflammatory response in conjunction with the activity of resident microbiota or invading pathogens, though the extent to which this occurs in human or environmental systems has yet to be quantified. Indeed, it has been observed that infected bacterial cells have manipulated and 'rewired' sulfur assimilation that will impact cysteine metabolism and likely sulfide production [70]. Furthermore, viruses encoding sulfur assimilation AMGs may be short-circuiting the assimilatory sulfur pathway by reducing the steps necessary for assimilation of sulfur into

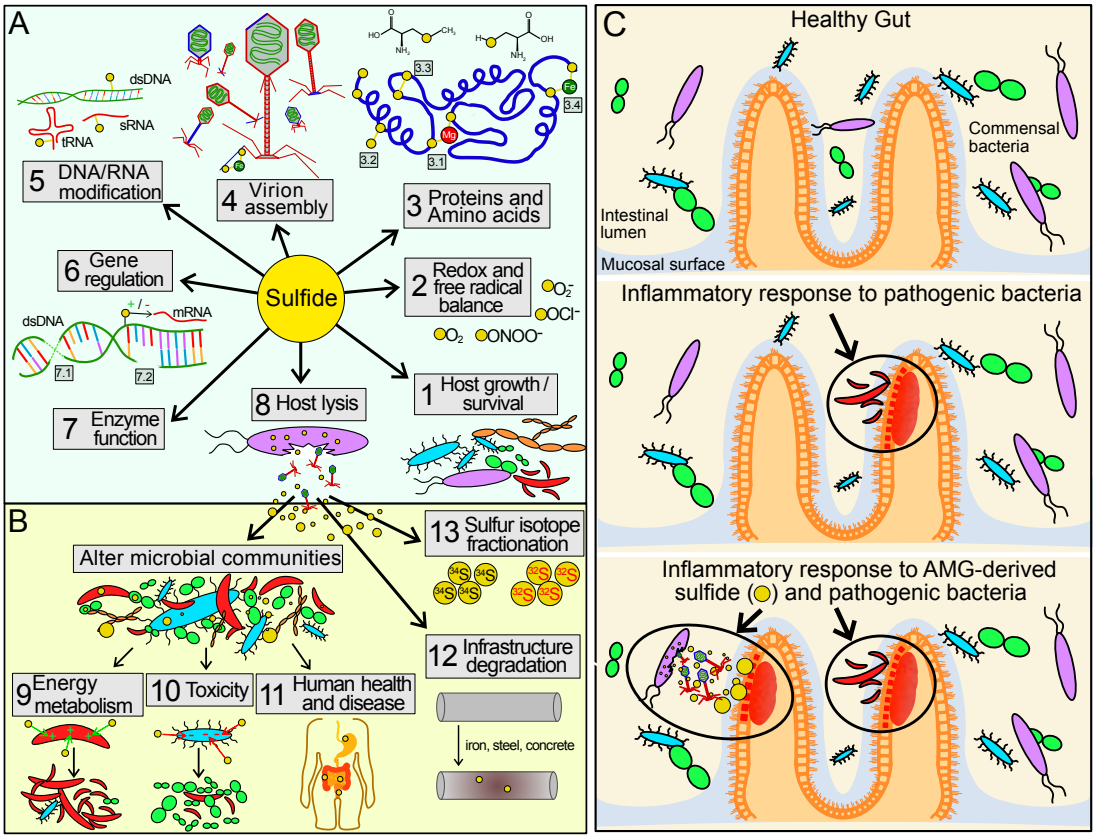


Figure 6. Virus-driven production of sulfide and its effects on human health, viral fitness and microbial communities. (A) Mechanisms by which sulfide could benefit viral fitness and (B) effect microbial communities, human health and environmental conditions. (C) Proposed impact of viral driven production of sulfide, in conjunction with activity of pathogenic bacteria, on inflammation in the gastrointestinal tract and its implications in IBD and CRC.

organosulfur compounds. This concept is supported by the observation that *cysH* is the most abundant organosulfur metabolism AMG, which plays a role in both the canonical sulfate assimilation pathway as well as direct sulfonation of organic molecules [182]. The latter mechanism may explain the high abundance of *cysH* on viral genomes.

The evidence presented here strongly points towards sulfide production as a component of viral organosulfur auxiliary metabolism, either directly or indirectly by AMG activity, which could provide many fitness advantages for viruses (**Figure 6A**). As obligate intracellular pathogens, viruses could benefit from the survival and enhanced growth of their host, which could be achieved by responding to sulfur starvation signals, assimilating sulfide for biosynthesis (e.g., for

sulfolipids), upregulating sulfide utilization (e.g., sulfide oxidation), antibiotic stress response (**Figure 6A.1**), or redox balance and free radical scavenging (**Figure 6A.2**) [35,53,82,135,183–185]. To benefit the virus directly, sulfide could be utilized for amino acid synthesis or protein function, such as for co-factor binding (e.g., metal ions) (**Figure 6A.3.1**), persulfidation of cysteine residues for signaling (**Figure 6A.3.2**), structural sulfide bridge formation (**Figure 6A.3.3**), iron-sulfur cluster formation (**Figure 6A.3.4**) or for viral structural proteins in virion assembly (**Figure 6A.4**) [186,187]. Furthermore, thiol modification of nucleic acids (i.e., dsDNA, tRNA and sRNA) could provide an avenue for responding to stresses (**Figure 6A.5**) or regulating gene expression for the virus or host (**Figure 6A.6**) [186,188–192]. Another method of nucleic acid modification that viruses may rely on is dsDNA recombination or integration (**Figure 6A.7.1**), or dsDNA repair (**Figure 6A.7.2**) which can be enabled by essential thiol components of enzymes [193–195]. Sulfide may even be a key component in the ability of viruses to effectively lyse their host (**Figure 6A.8**) [196].

However, due to the diversity of functions encoded by AMGs (e.g., degradation of organosulfur compounds directly into sulfide or sulfite, manipulation of organosulfur compound forms or fixing sulfur) it is likely that host physiology and local environmental conditions drive their acquisition and function. Regardless of the utility of AMGs employed by individual viruses, the eventual lysis and release of virus-derived sulfide or virus-influenced sulfur chemistry could have significant impacts on the surrounding environment and local microbial communities (**Figure 6B**). Increased sulfide concentrations could either enhance the growth of sulfide oxidizing organisms (**Figure 6B.9**) or act as a toxin to inhibit the growth of others (**Figure 6B.10**) [185]. Likewise, in both environmental and human systems, intracellular content released through viral lysis could alter nutrient availability and sulfide concentrations in the microbial community

(**Figure 6B.11**) or lead to the degradation of iron, steel and concrete in infrastructure (**Figure 6B.12**).

In humans, balancing organic and inorganic sulfur concentrations is pivotal to both the health of the gastrointestinal tract and the resident microbiota [197], and our evidence suggests that viruses may interfere with this equilibrium. Moreover, dozens of microbial species have been linked to accumulation of sulfide within the human gut via the degradation of organosulfur compounds (e.g., cysteine and taurine) and implicated in CRC and IBD [131,133], but the role of viruses in facilitating or upregulating these processes is unknown. Specifically, virus-mediated sulfide production could accelerate the development of sulfide-associated gastrointestinal disorders such as colitis, IBD and CRC (**Figure 6C**).

Our discovery of AMGs for organosulfur metabolism and sulfide production also has widespread ramifications for interpreting Earth history (**Figure 6B.13**). Sulfur isotope fractionation ($^{34}S/^{32}S$) analysis is widely used to interpret geological records and estimate rates of microbial processes such as sulfate reduction [198–200]. Microbial assimilatory sulfate reduction and viral auxiliary metabolism have been ignored as contributors to fractionation in the environment, mainly because sulfide is incorporated into organosulfur compounds instead of being exported into the environment as it is in dissimilatory reactions. As a result, assimilatory fractionation appears to be negligible (~3‰), whereas dissimilatory fractionation is frequently measured closer to 47‰ [201,202]. Without the incorporation of sulfide into organosulfur compounds, assimilatory sulfite to sulfide reduction fractionates up to 36-42‰ in *Salmonella*, *Clostridium* and *Bacillus* species [202]. We propose that virus-mediated sulfide production can directly impact the observed fraction of ^{32}S-enriched sulfide at scales relevant to dissimilatory sulfate reduction.

Overall, the global distribution and diversity of viruses encoding organosulfur transforming AMGs represents a so-far unexplored cog in the global organic and inorganic sulfur cycles. By modulating organic and inorganic sulfur compound concentrations, viruses likely play important roles in infrastructure degradation, human disease and ecosystem health. Beyond viral organosulfur metabolism, this study serves as a model for elucidating the impacts of virus-driven degradation of amino acids, whose fate is an important driver in human health and biotechnology and associated with ecosystem services in agriculture.

Limitations of Study

This study provides preliminary evidence for the function of organosulfur AMGs and viral influences on organosulfur compounds, namely hydrogen sulfide, in environmental and human microbiomes. One limitation is that the direct roles and interactions of AMGs within organosulfur metabolic frameworks and the elucidation of incurred benefits of hydrogen sulfide for some viruses was not shown. We show that the AMG *cysK* can be conserved evolutionarily over time which points towards, rather than measures, a fitness benefit of retaining the AMG. Furthermore, our experimental evidence for the benefit of sulfide to Lactococcus Phage P087, despite identifying viral CysK protein translation, did not distinguish if the measured fitness effect was the result of viral CysK or due to other unknown viral or host factors. Finally, our attempts to purify viral CysK protein and measure its activity in degrading cysteine was unsuccessful.

Table 1. Complete reaction(s) performed by each AMG-encoded protein. Each protein is grouped respective to the main organosulfur metabolism pathway in which it is involved. Full names of acronyms are as follows. PAP: adenosine 3',5'-bisphosphate, APS: adenosine 5'-phosphosulfate, PAPS: 3'-Phosphoadenosine-5'-phosphosulfate, CoA: Coenzyme A, OG: oxoglutarate, OAS: O-acetyl-L-serine, OASH: O-acetyl-L-homoserine, OSHS: O-succinyl-L-homoserine, SAMe: S-adenosyl-L-methionine, dAdoMT: S-adenosyl 3-

(methylsulfanyl)propylamine, MTA: S-methyl-5'-thioadenosine, MTR: 5-(methylsulfanyl)-α-D-ribose, MT: methylsulfanyl, SAH: S-adenosyl-L-homocysteine, SRH: S-ribosyl-L-homocysteine, DHK-MTPene: 1,2-dihydroxy-5-(methylsulfanyl)pent-1-en-3-one, ASSA: L-aspartate 4-semialdehyde, ASP: L-aspartyl-4-phosphate, MHO: L-methionine-(R)-S-oxide, Trdx: thioredoxin, THF: tetrahydrofolate, THP-3G: tetrahydropteroyl tri-L-glutamate.

Pathway	Protein	Reaction(s)
Assimilatory sulfate reduction	CysC	$APS + ATP \leftrightarrow PAPS + ADP + H^+$
	CysN	$SO_4^{2-} + ATP + H^+ \leftrightarrow APS + P_2O_7^{4-}$
	CysD	$SO_4^{2-} + ATP + H^+ \leftrightarrow APS^+ P_2O_7^{4-}$
	CysH	$PAP + SO_3^{2-} + \text{an oxidized Trdx} + 2\ H^+ \leftrightarrow PAPS + \text{a reduced Trdx}$
	CysNC	$PAP + ATP \leftrightarrow PAPS + ADP + H^+$
		$SO_4^{2-} + ATP + H^+ \leftrightarrow APS + P_2O_7^{4-}$
	CysJ	$SO_3^{2-} + 3\ NADPH + 5\ H^+ \rightarrow H_2S + 3\ NADP^+ + 3\ H_2O$
Direct sulfide production	CysK	$OAS + H_2S \rightarrow L\text{-cysteine} + acetate + H^+$
		$L\text{-cysteine} + H_2O \rightarrow pyruvate + H_2S + NH_4^+$
	CysM	$OAS + S_2O_3^{2-} \leftrightarrow S\text{-sulfo-L-cysteine} + acetate + H^+$
		$OAS + H_2S \rightarrow L\text{-cysteine} + acetate + H^+$
		$L\text{-cysteine} + H_2O \rightarrow pyruvate + H_2S + NH_4^+$
	MalY	$L\text{-cystathionine} + H_2O \rightarrow L\text{-homocysteine} + pyruvate + NH_4^+$
		$L\text{-cysteine} + H_2O \rightarrow pyruvate + H_2S + NH_4^+$
	DcyD	$D\text{-cysteine} + H_2O \rightarrow NH_4^+ + pyruvate + H_2S$
		$3\text{-chloro-D-alanine} + thioglycolate \rightarrow S\text{-carboxymethyl-D-cysteine} + Cl^- + H^+$
	MetC	$L\text{-cystathionine} + H_2O \rightarrow L\text{-homocysteine} + pyruvate + NH_4^+$
		$L\text{-cysteine} + H_2O \rightarrow pyruvate + H_2S + NH_4^+$
	MetY	$OASH + H_2S \leftrightarrow L\text{-homocysteine} + acetate + H^+$
Direct sulfite production	TauD	$taurine + 2\text{-OG} + O_2 \rightarrow SO_3^{2-} + 2\text{-aminoacetaldehyde} + succinate + CO_2 + H^+$
	SsuD	$\text{an alkylsulfonate} + FMNH_2 + O_2 \rightarrow \text{an aldehyde} + SO_3^{2-} + FMN + H_2O + 2H^+$
		$isethionate + FMNH_2 + O_2 \rightarrow glycolaldehyde + SO_3^{2-} + FMN + H_2O + 2H^+$
	MsmA	$methanesulfonate + NADH + O_2 \rightarrow formaldehyde + SO_3^{2-} + NAD^+ + H_2O$
Indirect sulfide production	MetB	$OSHS + L\text{-cysteine} \leftrightarrow L\text{-cystathionine} + succinate + H^+$
		$OSHS + H_2O \rightarrow 2\text{-oxobutanoate} + succinate + NH_4^+ + H^+$
	MetH	$L\text{-homocysteine} + \text{a 5-methyl-THF} \rightarrow L\text{-methionine} + \text{a THF}$
	MetE	$L\text{-homocysteine} + 5\text{-methyl-THP-3G} \leftrightarrow L\text{-methionine} + THP\text{-3G}$
	MetK	$ATP + L\text{-methionine} + H_2O \rightarrow SAMe + PO_4^{3-} + P_2O_7^{4-}$
	MtnN	$SAH + H_2O \rightarrow SRH + adenine$
		$MTA + H_2O \rightarrow MTR + adenine$
	Dcm	$SAMe + \text{a cytosine in DNA} \rightarrow \text{a 5-methylcytosine in DNA} + SAH + H^+$
	AhcY	$SAH + H_2O \rightarrow L\text{-homocysteine} + adenosine$
	LuxS	$SRH \rightarrow L\text{-homocysteine} + autoinducer\ 2$
	MsrC	$MHO + \text{a reduced Trdx} \rightarrow L\text{-methionine} + \text{an oxidized Trdx} + H_2O$
	MegL	$L\text{-methionine} + H_2O \rightarrow methanethiol + 2\text{-oxobutanoate} + NH_4^+$
	AspB	$L\text{-aspartate} + 2\text{-OG} \leftrightarrow oxaloacetate + L\text{-glutamate}$
		$L\text{-cysteine} + 2\text{-OG} \leftrightarrow 3\text{-mercaptopyruvate} + L\text{-glutamate}$
Indirect sulfur metabolism	CysE	$L\text{-serine} + acetyl\text{-CoA} \rightarrow OAS + CoA$
	NrnA	$PAP + H_2O \rightarrow AMP + PO_4^{3-}$
	SpeE	$putrescine + dAdoMT \leftrightarrow spermidine + MTA + H^+$
		$cadaverine + dAdoMT \rightarrow aminopropylcadaverine + MTA + H^+$
	MetA	$L\text{-homoserine} + succinyl\text{-CoA} \rightarrow OSHS + CoA$
	MtnK	$ATP + MTR \rightarrow ADP + 5\text{-MTR-1-phosphate} + H^+$
	MtnA	$5\text{-MTR-1-phosphate} \rightarrow 5\text{-(MT)-ribulose 1-phosphate}$
	MtnD	$DHK\text{-MTPene} + O_2 \rightarrow 4\text{-(MT)-2-oxobutanoate} + formate + H^+$
		$DHK\text{-MTPene} + O_2 \rightarrow 3\text{-(MT)propanoate} + formate + CO + H^+$

	LysC	L-aspartate + ATP → ASP + ADP
	ThrA	L-aspartate + ATP → ASP + ADP
		ASSA + NAD(P)H + H^+ → L-homoserine + $NAD(P)^+$
	Asd	ASSA + $NADP^+$ + PO_4^{3-} ↔ ASP + NADPH + H^+
	Hom	ASSA + NAD(P)H + H^+ → L-homoserine + $NAD(P)^+$
	Mdh	(S)-malate + NAD^+ ↔ oxaloacetate + NADH + H^+
	CysQ	PAP + H_2O → AMP + PO_4^{3-}

Methods

RESOURCE AVAILABILITY

Materials Availability

The recombinant phage line generated in this study is available upon request.

Data and Code Availability

All sequences used in this study are publicly available and can be found at their original sources. The genomic and protein sequences of viruses highlighted in this study and respective AMG protein sequences identified can be found on GitHub (https://github.com/AnantharamanLab/Kieft_et_al_2021_organosulfur_AMGs) and Zenodo (http://doi.org/10.5281/zenodo.4947151). This paper does not report original code. Any additional information required to analyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Lactococcus system growth conditions

Lactococcus lactis subs. lactis C10 and Lactococcus phage P087 were obtained from Université Laval's Félix d'Hérelle Reference Center for Bacterial Viruses (Canada, www.phage.ulaval.ca). *L. lactis* C10 was grown without agitation at 30°C in M17 broth (Oxoid) supplemented with 0.5%

glucose (GM17). Infections were supplemented with 10mM $CaCl_2$ and incubated without agitation at room temperature.

T7 system growth conditions

T7 phage was obtained from ATCC (ATCC® BAA-1025-B2). *Saccharomyces cerevisiae* BY4741 and *E. coli* BL21 are lab stocks, and *E. coli* 10G is a highly competent DH10B derivative [203] originally obtained from Lucigen (60107-1). *E. coli* BW25113 and BW25113Δ*cysK* were obtained from Doug Weibel (University of Wisconsin, Madison).

All bacterial hosts were grown in and plated on LB media (1% Tryptone, 0.5% Yeast Extract, 1% NaCl in dH_2O, plates additionally contain 1.5% agar, while top agar contained 0.5% agar) and LB media was used for all experimentation. All incubations of bacterial cultures were performed at 37°C, with liquid cultures shaking at 200-250 rpm unless otherwise specified. Bacterial hosts were streaked on appropriate LB plates and stored at 4°C. *S. cerevisiae* BY4741 was grown on YPD (2% peptone, 1% yeast extract, 2% glucose in dH_2O, plates additionally contain 2.4% agar), after Yeast Artificial Chromosomes (YAC) transformation *S. cerevisiae* BY4741 was grown on SD-Leu (0.17% yeast nitrogen base, 0.5% ammonium sulfate, 0.162% amino acids – Leucine [Sigma Y1376], 2% glucose in dH_20, plates additionally contain 2% agar). All incubations of *S. cerevisiae* were performed at 30°C, with liquid cultures shaking at 200-250 rpm. *S. cerevisiae* BY4741 was streaked on YPD or SD-Leu plates as appropriate and stored at 4°C.

T7 phage was propagated using *E. coli* BL21 after initial receipt from ATCC and then as described on various hosts in methods. All phage experiments were performing using LB and culture conditions as described for bacterial hosts. Phages were stored in LB at 4°C. For long term

storage all microbes were stored as liquid samples at -80°C in 10% glycerol, 90% relevant media. SOC (2% tryptone, 0.5% yeast extract, 0.2% 5M NaCl, 0.25% 1M KCL, 1% 1M $MgCl_2$, 1% 1M $MgSO_4$, 2% 1M glucose in dH_2O) was used to recover host and phages after transformation.

For infection experiments, stationary phase cultures were created by growing bacteria overnight (totaling ~20-30 hours of incubation) at 37°C. Exponential phase culture consisted of stationary culture diluted 1:20 in LB then incubated at 37°C until an OD_{600} of ~0.4-0.8 was reached, typically after 40 minutes. Phage lysate was purified by centrifuging phage lysate at 16g, then filtering supernatant through a 0.22 μM filter. To establish titer, phage samples were serially diluted (1:10 or 1:100 dilutions made to 1 mL in 1.5 mL microcentrifuge tubes) in LB to a 10^{-8} dilution for titering by spot assay. Spot assays were performed by mixing 250 μl of relevant bacterial host in the stationary phase with 3.5 mL of 0.5% top agar, briefly vortexing, then plating on LB plates pre-warmed to 37°C. After plates solidified (typically ~5 minutes), 1.5 μl of each dilution of phage sample was spotted in series on the plate. Plates were incubated and checked every 2-4 hours or overnight (~20-30 hours) to establish a preliminary titer. MOI was estimated by calculated by dividing phage titer by estimated bacterial concentration.

METHOD DETAILS

Identification of viral genomes

A total of 125,842 viral genomes from the Integrated Microbial Genomes/Virus (IMG/VR) [141] v1 database were used for analysis (accessed October 2017). Only publicly available genomes >5kb analyzed by Paez-Espino *et al*. (2016) were used in this study [9]. Open reading frames were predicted using Prodigal with default parameters (v2.6.3) [98]. All viral genomes were annotated using a combination of Prokka (v1.13.3) [204], Integrated Microbial Genomes and Microbiomes

pipeline [205], and InterProScan (v65.0) [206]. Contigs with a high ratio of bacterial to viral protein annotations were manually identified and discarded. Contigs were further validated and annotated using a combination of VIBRANT (v1.2.1) and VirSorter (v1.0.3, virome database, categories 1, 2, 4, 5) [115,117]. All viral genomes encoding AMGs were manually inspected. Additional viral genomes were identified on the National Center for Biotechnology Information (NCBI) RefSeq [207–209] or Genbank database [210] (accessed Jan 2019) by querying viral genomes for AMGs of interest by blastp domain analysis [101,211]. Approximately 9,500 genomes corresponding to the viral classification *Caudovirales* were searched. VIBRANT and VirSorter were used to identify viruses >5kb from Lake Mendota, WI.

AMG identification and annotation

In-house hidden Markov model (HMM) profiles were built corresponding to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway of organosulfur Metabolism as well as Cysteine and Methionine Metabolism (accessed December 2018) [212]. The two pathways' KEGG Orthology (KO) numbers (189 total) were used to access corresponding proteins from the UniProt database (release 2018_11) [213]. The resulting proteins were aligned with MAFFT (v7.388, default parameters) [214] and HMM profiles were built using hmmbuild (HMMER v3.1, default parameters) [102]. HMM profiles for CysC and CysH were built in the same manner, except manually verified viral CysC and CysH sequences, respectively, were added to the alignment for robustness. Hmmsearch (HMMER v3.1, evalue < 1e-5) was used to scan proteins on viral genomes. Proteins identified by the in-house HMM profiles were uploaded to the KEGG BlastKOALA server (v2.1) [215] and queried under "prokaryotes" taxonomy and "genus prokaryotes" database for best hit annotations. Proteins annotated according to the original 189 KO numbers were selected for further

verification. Manual verification of several representatives of each KO number (i.e., protein family) was done to curate the results using blastp (NCBI non-redundant database, accessed Jan 2019) and InterProScan (v71.0) to check for the presence of all expected conserved domains. Individual proteins and protein families of irrelevance or incorrect annotation were removed.

Sequence alignment and dN/dS analysis

Alignment of CysH, CysK, CysC, TauD and MetK sequences was performed using MAFFT (v7.388, default parameters). For *cysH*-encoding genomes identified from NCBI, all viral sequences were used. Host genomes were scanned, by annotation and blastp domain analysis, for multiple copies of *cysH* and all those identified were used, along with non-host bacterial sequences that were found to be highly similar to viral sequences according to pairwise identity. For the remaining alignments, all viral AMG protein sequences that shared at least 95% pairwise identity were restricted to one representative using CD-HIT (accessed Jan 2019) [216–218] and aligned. Viral CysK and CysH sequences were limited to lengths 200-330 and 117-600 amino acids, respectively. To obtain bacterial representatives, the majority consensus sequence of aligned viral proteins was queried against the NCBI RefSeq database by blastp (evalue < 1e-5). In order to ensure broad phylogenetic distribution of blastp results, the output was restricted to the top 500 hits from each of five phylogenetic groups based on NCBI categorization: [1] Proteobacteria, [2] Terrabacteria, [3] FCB superphlylum, [4] PVC superphylum and [5] a group containing all other phyla. The resulting sequences were manually limited to specific lengths to match viral sequences (CysC: 210-360, CysH: 150-600, CysK: 269-400, TauD: 314-400 amino acids, MetK: all) and reduced to one representative per 50% pairwise identity using CD-HIT. Viral and bacterial representatives were aligned together using MAFFT (default parameters) and gaps were stripped by 98%. The

resulting alignments were used for phylogenetic analysis. Visualization of alignments was done using Geneious Prime 2019.0.3. For reference to full virus protein name and genome, see Table S1.

The AMGs for *cysK, cysC, cysD, cysH, tauD, msmA, metK, mtmN and luxS* were used to calculate *dN/dS* ratios. dRep (v2.6.2) was used to compare AMG sequences separately (dRep compare --SkipMash --S_algorithm goANI) and dnds_from_drep.py was used to calculate *dN/dS* ratios from the AMG pairs [219]. The *dN/dS* ratios were visualized with Seaborn (v0.8.1) and Matplotlib (v3.0.0).

Sequence phylogeny

Phylogenetic analysis was performed using protein alignments of CysH, CysK, CysC, TauD and MetK as described above. To infer phylogenetic relationships RAxML (v8.2.4) [220] was used with the following parameters: raxmlHPC-PTHREADS -N 100 -f a -m PROTCATLG. Resulting best trees were used and rooted by manual identification of most distant (outgroup) taxa. Trees were visualized using FigTree (v1.4.3) [221].

Protein functional analysis

For domain and residue analysis, phylogenetic trees were used as a reference to select representative viral and bacterial sequences, which were then aligned using MAFFT (default parameters). Annotations of functional amino acid residues were labeled according to the Protein Data Bank (PDB, accessed January 2019)[222] with the following identification numbers: 4BZQ and 4BZP (CysC), 2GOY (CysH), 3ZEI (CysK), 3SWT (TauD), and 1RG9 (MetK). For alignments with no phylogenetic tree, up to five viral sequences and five PDB homologs (when available)

were randomly selected for all AMGs with abundance of five or greater. The PDB sequences used for annotation were added to the alignment. N- and C-terminal ends of protein alignments were manually removed for clarity and gaps were stripped by 90% (for alignments with phylogenetic trees) or 80% (for all others). Residues were highlighted according to 85% pairwise identity between sequences, excluding sequence gaps. An identity graph, generated by Geneious, was fitted to the alignment to visualize pairwise identity of 100% (green), 99-30% (yellow) and 29-0% (red).

Protein Reactions

Enzymatic reactions, diagrams and pathways were created by referencing KEGG and MetaCyc (v22.6) [223] annotations.

Viral transcriptomics and growth rates

Publicly available metatranscriptomic data from Lake Mendota, WI was assessed for AMGs by querying annotation names [224]. This gene expression data comprises a two-day time series and is accompanied by metagenomic assemblies (IMG Taxon Object IDs 3300013004 and 3300013005). Metatranscriptomic reads were mapped to a custom, non-redundant database of freshwater reference data, including the metagenome assemblies; annotations in this study are derived from the annotations of the reference database. We used read counts normalized to transcripts per liter as the input for our study, and we searched for AMGs in the metagenomic assemblies as described above.

The growth rate of the *cysC*-encoding Lake Mendota virus was identified using index of replication (iRep) with default parameters [156]. Metagenomic assembly reads used for iRep are available on IMG under the Taxon Object ID 3300013005. Reads were mapped to the viral genome

using Bowtie2 (v2.3.4.1) [225]. GC-skew to indicate rolling circle replication of the viral genome was likewise completed using the iRep toolkit.

Virus growth and fitness assay

Approximately 10^8 plaque forming units (PFUs) of Lactococcus phage P087 (approximate multiplicity of infection (MOI) of 1 were used to infect 1mL of *L. lactis* C10 which had been brought to an optical density (OD_{600}) of approximately 0.15 in GM17 broth. For fitness experiments, either vehicle control (water), 10µM Na_2S or 100µM Na_2S was supplemented to the media at time of infection. Infections were incubated without agitation at room temperature for approximately three hours. Additional cultures of uninfected *L. lactis* C10 with all other variables identical were measured for growth at the endpoint of infections using OD_{600}. To end infections, *L. lactis* C10 were spun out of solution at 10,000 rcf and the supernatant (i.e viral fraction) was removed and cooled to 4°C. Plaque assays were done using the standard double agar method [226] with diluted viral fraction and *L. lactis* C10 brought to high concentration. A 1% bottom agar and 0.4% top agar of GM17 were used, both supplemented with 0.5% glycine and 10mM $CaCl_2$.

Virus and host cysK qPCR assay

An overnight culture of *L. lactis* C10 was diluted in GM17 broth to OD 0.08 and grown at 30°C for ~2 hours until OD reached 0.15. In a batch culture 10mM $CaCl_2$ was added. Two different conditions were assayed, each in duplicate (biological replicates): (1) *L. lactis* C10 control and (2) *L. lactis* C10 plus Lactococcus phage P087. For infection conditions, Lactococcus phage P087 was added at a MOI of 1 (time 0 minutes). RNA was extracted using the PureLink RNA Mini Kit (Ambion) from 500µL of the cellular fraction at 15, 60 and 120 minutes post-infection. RNA was

then treated with DNase with the DNase Max Kit (Qiagen) and converted to cDNA using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). qPCR of viral and host *cysK* was performed using *Power* SYBR Green PCR Master Mix (Applied Biosystems) with 7ng of cDNA template and the following primer sets (IDT): *L. lactis* C10 forward (CCTTCGTTGGCTCTGCTTTG), *L. lactis* C10 reverse (TGGCATCATCTCCTTTGACCC), Lactococcus phage P087 forward (CAGAAACTATCGGAAACACACCAC), and Lactococcus phage P087 reverse (TTGAGTGAATGACCTGCTCCA) (**Table S10**). The concentration of template cDNA was measured with the Qubit dsDNA BR Assay Kit (Invitrogen). The viral and host *cysK* sequences were sufficiently dissimilar in sequence identity (<60% at the protein level) to allow for accurate distinction by qPCR and the primers selected.

Mass spectrometry and protein identification

L. lactis C10 was grown without agitation at 30°C in modified M17 broth supplemented with 0.5% glucose (mGM17). mGM17 was made by adding 1.25g glucose, 0.625g tryptone, 1.25g peptone, 0.125g yeast extract, 0.125g ascorbic acid, 0.0626g anhydrous magnesium sulfate and 4.75g disodium glycerophosphate to 250mL deionized water. Approximately 10^8 PFUs of Lactococcus phage P087 were used to infect 3mL of *L. lactis* C10 which had been brought to OD_{600} of approximately 0.15 and supplemented with 10mM $CaCl_2$. Infections proceeded to complete lysis without agitation at room temperature for approximately three hours. To end the infection, *L. lactis* C10 were spun out of solution at 10,000 rcf and the supernatant was removed and stored at 4°C. The supernatant was size fractionated by filtration for the 100kDa to 10kDa size fraction before trypsin solution digestion and analysis by Long Orbitrap LC/MS/MS (University of Wisconsin-Madison Biotechnology Center).

Genome organization and comparisons

Genome organization was visualized using Geneious Prime. Genes were manually colored by referencing functions according to NCBI RefSeq or Genbank annotation, or blastp search. Viral genomes in genbank format were compared and visualized with EasyFig (v2.2.2) [227] using the tblastx function. Only tblastx (v2.8.1+) hits with percent identities greater than 30% and e-values less than 0.001 are shown. Remaining analysis parameters were set to default. Circular sequences were visualized linearly for ease of comparison.

Geographical distributions

IMG Taxon Object ID numbers were used to identify global coordinates of studies in which AMGs were identified. Coordinates were mapped using Matplotlib's Basemap (v1.2.0) [228]. Human studies were excluded from coordinate maps.

Host classification

GhostKOALA (v2.0) [215] with the "genus prokaryotes" database was used to query all 3,794 AMG-encoded proteins identified from IMG/VR derived viruses (3,421 annotated and used for taxonomy). To benchmark accuracy of the analysis, all 282 AMG-encoded proteins identified from NCBI-derived viruses with known hosts were queried in the same manner (278 were annotated and used for taxonomy) and compared to the taxonomy of hosts.

T7 recombination: cloning

All primers can be found in **Table S10**. PCR was performed using KAPA HiFi (Roche) for all experiments with the exception of multiplex PCR for screening Yeast Artificial Chromosomes (YACs), which was performed using KAPA2G Robust PCR kits (Roche). DNA purification was performed using EZNA Cycle Pure Kits (Omega Bio-tek) using the centrifugation protocol. YAC extraction was performed using YeaStar Genomic DNA Extraction kits (Zymo Research). All cloning was performed according to manufacturer documentation except where noted in methods. PCR reactions using phage as template use 1 µl of undiluted phage stock, with extension of the 95°C denaturation step to 5 minutes.

Electroporation of YACs was performed using a Bio-rad MicroPulser (165-2100), Ec2 setting (2 mm cuvette, 2.5 kV, 1 pulse) using 50 µl competent cells and 2 µl YAC DNA for transformation. Electroporated cells were immediately recovered with 950 µl SOC, then incubated at 37°C for 1 to 1.5 hours and plated or grown in Lb.

E. coli 10G competent cells were made by adding 8 mL overnight 10G cells to 192 mL SOC (with antibiotics as necessary) and incubating at 21°C and 200 rpm until ~OD_{600} of 0.4 as determined using an Agilent Cary 60 UV-Vis Spectrometer using manufacturer documentation (actual incubation time varies based on antibiotic, typically overnight). Cells are centrifuged at 4°C, 800-1000g for 20 minutes, the supernatant is discarded, and cells are resuspended in 50 mL 10% glycerol. Centrifugation and washing are repeated three times, then cells are resuspended in a final volume of ~1 mL 10% glycerol and are aliquoted and stored at -80°C. Cells are competent for plasmid and YACs. All primers used in experiments in this publication are listed in supplemental.

T7 recombination: engineering T7 with cysK

Phages were assembled using YAC rebooting [229,230], which requires yeast transformation of relevant DNA segments, created as follows. A prs415 yeast centromere plasmid was split into three segments by PCR, separating the centromere and leucine selection marker, which partially limits recircularization and improved assembly efficiency [231]. Wildtype T7 segments were made by PCR using wildtype T7 as template. *CysK* segments were made by colony PCR of BW25113. *CysK* was inserted into two locations to create two phage constructs. The first location was replacement of gp1.7 to establish *CysK* in early Class II genes. This insertion causes a two amino acid extension (YE) of the immediate 5' gene gp1.6 that was not anticipated to have an effect on phage viability. The second location was inserted adjacent to gp6.3 to establish *CysK* in early class III genes and leverages a copy of phage promoter phi6.5 for expression.

DNA parts were combined together (0.1 pmol/segment) and transformed into *S. cerevisiae* BY4741 using a high efficiency yeast transformation protocol [232] using SD-Leu selection. After 2-3 days colonies were picked and directly assayed by multiplex colony PCR to assay assembly. Multiplex PCR interrogated junctions in the YAC construct and was an effective way of distinguishing correctly assembled YACs. Correctly assembled YACs were purified and transformed into *E. coli* 10G cells and these cultures incubated until lysis, after which phages were purified to create the initial phage stock.

T7 recombination: passaging and AMG retention

Either T7Δ1.7::cysK or T7::cysK phages were added to 5 mL exponential phase BW25113 or BW25113Δ*cysK* at an estimated MOI of 10^{-4} to allow for an estimated three phage passages. After the culture had fully lysed, typically ~1 hour and 30 minutes, lysate was purified and then the titer established by spot assay. This process was then repeated twice for a total of an estimated

9 phage passages assuming at least 100 phage progeny per host. Phage lysate from the final passage was used as template for sequencing to determine if the *cysK* insert remained as the consensus sequence in the phage population. The entire process was repeated in biological triplicate for both host and phage combinations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Virus growth and fitness

The number of resulting plaques from the growth and fitness assays were normalized to 100% of controls for each experiment. Three independent experiments with three biological infection replicates and two biological growth replicates each was performed. Further information of experiments can be found in Method Details below.

Virus and host cysK qPCR

For each replicate of the two conditions assayed both primer sets were used for qPCR. To analyze the qPCR results, the Cq readings were averaged between the three replicates for each treatment at each timepoint to obtain a single datapoint per treatment:primer pair per timepoint, termed *average Cq*. Using time point zero for the uninfected *L. lactis* C10 condition with *L. lactis* C10 *cysK* primers as the baseline *control*, delta-delta-Cq values were calculated by subtracting the *control* value from the *average Cq* values. This result calculates the expression of *L. lactis* C10 *cysK* at time point zero to be normalized to zero (delta-delta-Cq of zero). Finally, all delta-delta-Cq values were transformed using the formula $2^{-(\text{delta-delta-Cq})}$ [233]. All raw Cq values and normalized results, including equations, can be found in **Table S6**. Further information of experiments can be found in Method Details below.

Acknowledgments

We thank Anna-Louise Reysenbach and Katherine D. McMahon for helpful discussions and suggestions. We thank the University of Wisconsin - Office of the Vice Chancellor for Research and Graduate Education, University of Wisconsin – Department of Bacteriology, and University of Wisconsin – College of Agriculture and Life Sciences for their support. K.K. is supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison. This work was partly supported by the National Science Foundation grant OCE-2049478 to K.A. A.J.P. was supported by the Ministry of Culture and Science of North Rhine-Westphalia (Nachwuchsgruppe "Dr. Alexander Probst") and the NOVAC project of the German Science Foundation (grant number DFG PR1603/2-1). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

Author contributions

K.K. and K.A. designed the study. K.K., E.Z., P.H., and A.M.B. performed host-virus experiments. K.K. and K.A. conducted bioinformatic and metabolic analyses. K.K. and A.L. performed metatranscriptomic analyses. K.K. and K.A drafted the manuscript. All authors (K.K., A.M.B., P.H., A.M.L., E.Z., Z.Z., J.R., S.P.E., A.J.P., S.R., S.R., and K.A.) reviewed the results and approved the manuscript.

Chapter 3: Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages

Kristopher Kieft[1#], Zhichao Zhou[1#], Rika E. Anderson[2], Alison Buchan[3], Barbara J. Campbell[4], Steven J. Hallam[5,6,7,8,9], Matthias Hess[10], Matthew B. Sullivan[11], David A. Walsh[12], Simon Roux[13], Karthik Anantharaman[1]

[1] Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, 53706, USA
[2] Biology Department, Carleton College, Northfield, Minnesota, USA
[3] Department of Microbiology, University of Tennessee, Knoxville, TN, 37996, USA
[4] Department of Biological Sciences, Life Science Facility, Clemson University, Clemson, SC, 29634, USA
[5] Department of Microbiology & Immunology, University of British Columbia, Vancouver, British Columbia V6T 1Z3, Canada
[6] Graduate Program in Bioinformatics, University of British Columbia, Genome Sciences Centre, Vancouver, British Columbia V5Z 4S6, Canada
[7] Genome Science and Technology Program, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[8] Life Sciences Institute, University of British Columbia, Vancouver, British Columbia, Canada
[9] ECOSCOPE Training Program, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3
[10] Department of Animal Science, University of California Davis, Davis, CA, 95616, USA
[11] Department of Microbiology, The Ohio State University, Columbus, OH, 43210, USA
[12] Groupe de recherche interuniversitaire en limnologie, Department of Biology, Concordia University, Montréal, QC, H4B 1R6, Canada
[13] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[#]These authors contributed equally

Publication:
Kieft, K. and Zhou, Z. *et al.* Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat Commun* **12**, 3503 (2021).

All supplementary figures, tables, and files are available at the following Figshare repository:
https://figshare.com/projects/Kristopher_Kieft_PhD_Dissertation/136427

Abstract

Microbial sulfur metabolism contributes to biogeochemical cycling on global scales. Sulfur metabolizing microbes are infected by phages that can encode auxiliary metabolic genes (AMGs) to alter sulfur metabolism within host cells but remain poorly characterized. Here we identified 191 phages derived from twelve environments that encoded 227 AMGs for oxidation of sulfur and thiosulfate (*dsrA*, *dsrC/tusE*, *soxC*, *soxD* and *soxYZ*). Evidence for retention of AMGs during niche-differentiation of diverse phage populations provided evidence that auxiliary metabolism imparts measurable fitness benefits to phages with ramifications for ecosystem biogeochemistry. Gene abundance and expression profiles of AMGs suggested significant contributions by phages to sulfur and thiosulfate oxidation in freshwater lakes and oceans, and a sensitive response to changing sulfur concentrations in hydrothermal environments. Overall, our study provides fundamental insights on the distribution, diversity, and ecology of phage auxiliary metabolism associated with sulfur and reinforces the necessity of incorporating viral contributions into biogeochemical configurations.

Introduction

Viruses that infect bacteria (bacteriophages, or phages) are estimated to encode a larger repertoire of genetic capabilities than their bacterial hosts and are prolific at transferring genes throughout microbial communities[41,234–236]. The majority of known phages have evolved compact genomes by minimizing non-coding regions, reducing the average length of encoded proteins, fusing proteins and retaining few non-essential genes[109,237]. Despite their reduced genome size and limited coding capacity, phages are known for their ability to modulate host cells during infection, take over cellular metabolic processes and proliferate through a bacterial population, typically

through lysis of host cells[36,38]. Phage-infected hosts, termed virocells, take on a distinct physiology compared to an uninfected state[70]. According to some estimates, as many as 20-40% of all bacteria in aquatic environments are assumed to be in a virocell state, undergoing phage-directed metabolism[37,57]. This has led to substantial interest in understanding the mechanisms that provide phages with the ability to redirect nutrients within a host and ultimately how this manipulation may affect microbiomes and ecosystems.

One such mechanism by which phages can alter the metabolic state of their host is through the activity of phage-encoded auxiliary metabolic genes (AMGs)[73,74]. AMGs are typically acquired from the host cell (i.e., recombined onto the phage genome) and can be utilized during infection to augment or redirect specific metabolic processes within the host cell[44,75,140]. These augmentations likely function to maintain, drive or short-circuit important steps of a metabolic pathway and can provide the phage with sufficient fitness advantages under specific metabolic or nutrient conditions in order to retain these genes over time[74,76]. Two notable examples of AMGs are core photosystem II proteins *psbA* and *psbD*, which are commonly encoded by phages infecting Cyanobacteria in both freshwater and marine environments, and responsible for supplementing photosystem function in virocells during infection[77,78,238,239]. PsbA and PsbD play important roles in maintenance of photosynthetic energy production over time within the host; this energy is subsequently utilized for the production of resources (e.g., nucleotides) for phage propagation[74,75]. The Cyanobacteria host does not benefit from the additional gene copy (i.e., phage AMG) since the replication benefits and energy acquisition are in favor of the infecting phage. Other descriptions of AMGs include those for sulfur oxidation in the pelagic oceans[82,140], methane oxidation in freshwater lakes[81], ammonia oxidation in surface oceans[79], carbon utilization (e.g., carbohydrate hydrolysis) in soils[84,85], and marine ammonification[240]. As a further example, it has

been hypothesized that some phages encoding carbon utilization AMGs function to redirect carbon from glycolysis to dNTP synthesis, for phage genome replication, by inducing a state of host starvation[36]. In this scenario, the phages encode their own AMG for specific manipulation of host processes rather than simply providing an extra gene copy to the host. Beyond these examples, the combined effect of phage auxiliary metabolism on ecosystems scales has yet to be fully explored or implemented into conceptualizations of microbial community functions and interactions.

Dissimilatory sulfur metabolism (DSM) encompasses both reduction (e.g., sulfate to sulfide) and oxidation (e.g., sulfide or thiosulfate to sulfate) and accounts for the majority of sulfur metabolism on Earth[128]. Bacteria capable of DSM (termed as sulfur microbes) are phylogenetically diverse, spanning 13 separate phyla, and can be identified throughout a range of natural and human systems, aquatic and terrestrial biomes, aerobic or anaerobic environments, and in the light or dark[132]. Since DSM is often coupled with primary production and the turnover of buried organic carbon, understanding these processes is essential for interpreting the biogeochemical significance of both microbial- and phage-mediated nutrient and energy transformations[132]. Phages of DSM-mediating microorganisms are not well characterized beyond the descriptions of phages encoding *dsrA* and *dsrC* genes infecting known sulfur oxidizers from the SUP05 group of Gammaproteobacteria[82,140], and viruses encoding *dsrC* and *soxYZ* genes associated with proteobacterial hosts in the epipelagic ocean[35]. Despite the identification of DSM AMGs across multiple host groups and environments, there remains little context for their global diversity and roles in the biogeochemical cycling of sulfur. Characterizing the ecology, function and roles of phages associated with DSM is crucial to an integral understanding of the mechanisms by which sulfur species are transformed and metabolized.

Here we leveraged publicly available metagenomic and metatranscriptomic data to identify phages capable of manipulating DSM within host cells. We identified 191 phages encoding AMGs for oxidation and disproportionation of reduced sulfur species, such as elemental sulfur and thiosulfate, in coastal ocean, pelagic ocean, hydrothermal vent, human, and terrestrial environments. We refer to these phages encoding AMGs for DSM as *sulfur phages*. These sulfur phages represent different taxonomic clades of *Caudovirales*, namely from the families *Siphoviridae*, *Myoviridae* and *Podoviridae*, with diverse gene contents, and evolutionary history. Using paired viral-host gene coverage measurements from metagenomes recovered from hydrothermal environments, freshwater lakes, and *Tara* Ocean samples, we provide evidence for the significant contribution of viral AMGs to sulfur and thiosulfate oxidation. Investigation of metatranscriptomic data suggested that phage-directed sulfur oxidation activities showed significant increases with the increased substrate supplies in hydrothermal ecosystems, which indicates rapid and sensitive responses of virocells to altered environmental conditions. Overall, our study provides key insights on the distribution, diversity, and ecology of phage-directed dissimilatory sulfur and thiosulfate metabolisms and reinforces the need to incorporate viral contributions into assessments of biogeochemical cycling.

Results

Unique sulfur phages encode AMGs for oxidation of elemental sulfur and thiosulfate

We queried the Integrated Microbial Genomes/Viruses (IMG/VR v2.1) database for phages encoding genes associated with pathways for dissimilatory sulfur oxidation and reduction processes. We identified 190 metagenomic viral contigs (mVCs) and one viral single-amplified genome[241] carrying genes encoding for reverse dissimilatory sulfite reductase subunits A and C

(*dsrA* and *dsrC*), thiouridine synthase subunit E (*tusE*, a homolog of *dsrC*), sulfane dehydrogenase subunits C and D (*soxC*, *soxD*), and fused sulfur carrier proteins Y and Z for thiosulfate oxidation (*soxYZ*). All mVCs except one (KiloMoana_10000689) were estimated to be partial genome scaffolds. While phages carrying *dsrA*, *dsrC*/*tusE* and *soxYZ* have been previously described in specific marine environments, this is the first report of *soxC* and *soxD* encoded on viral genomes. Each identified mVC encoded between one to four total DSM AMGs for a total of 227 AMGs (Fig. 1a, Supplementary Data 1). The mVCs ranged in length from 5 kb to 308 kb, with an

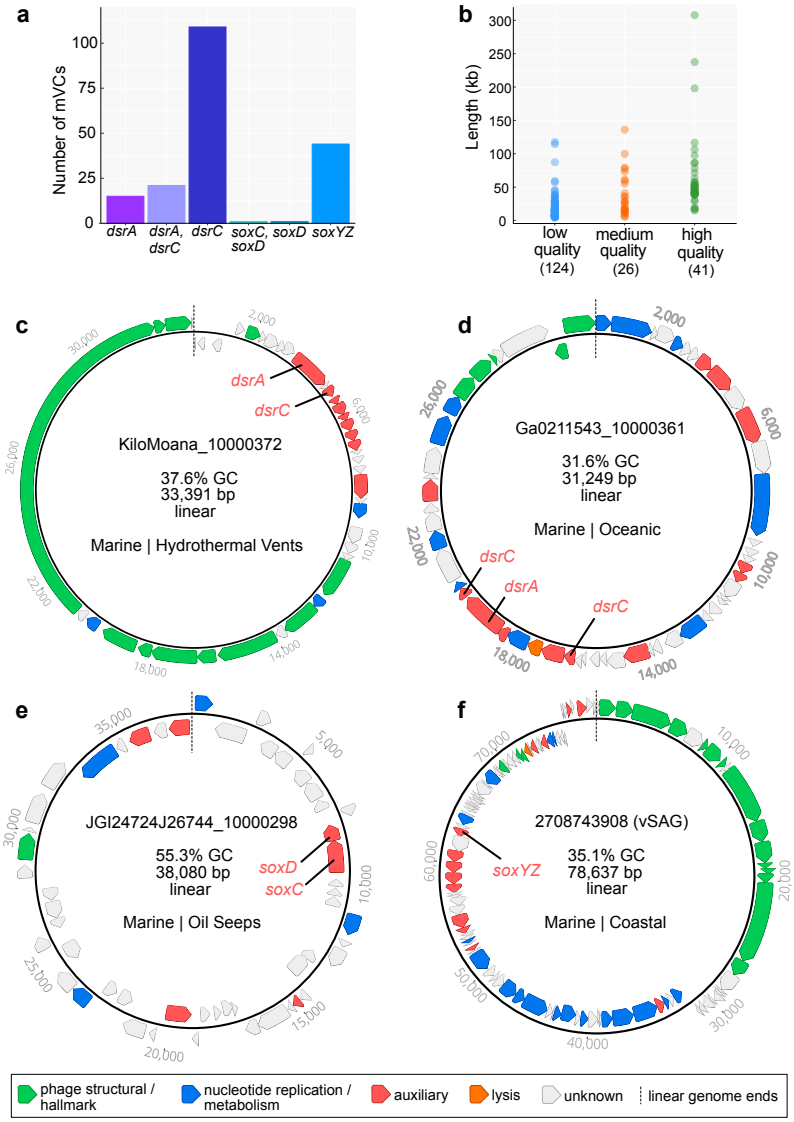


Fig. 1 Dataset summary statistics and representative genome organization diagrams of mVCs. **a** The number of mVCs, 191 total, encoding single or multiple DSM AMGs. **b** Estimated mVC genome qualities as a function of scaffold lengths. mVCs encoding **c** *dsrA* and *dsrC*, **d** *dsrA* and two *dsrC*, **e** *soxC* and *soxD*, and **f** *soxYZ*. For **c**, **d**, **e** and **f** linear mVC scaffolds are visualized as circular with the endpoints indicated by dashed lines, and predicted open reading frames are colored according to VIBRANT annotation functions. Abbreviation: vSAG, viral single-amplified genome; GC, guanine-cytosine content; bp, base pairs.

average length of approximately 31 kb and a total of 83 sequences greater than 20 kb. The mVCs consisted of 124 low-, 26 medium- and 41 high-quality draft scaffolds according to quality estimations based on gene content (Fig. 1b). Only one mVC was a complete circular genome and was identified as previously described[82]. The majority of viruses in this study, with the exception of several mVCs encoding *tusE*-like AMGs were predicted to have an obligate lytic lifestyle on the basis of encoded proteins functions.

The mVCs displayed unique and diverse genomic arrangements, regardless of the encoded AMG(s). However, in most cases the encoded AMGs were found within auxiliary gene cassettes, separate from structural and nucleotide metabolism cassettes (Fig. 1c, d, e, f). Auxiliary cassettes in phages typically encode genes that are not essential for productive propagation but can provide selective advantages during infection, such as in specific nutrient limiting conditions or to overcome metabolic bottlenecks[242]. This genomic arrangement suggests that the role of DSM AMGs is related to host modulation rather than essential tasks such as transcription/translation, genome replication or structural assembly.

Validation of conserved amino acid residues and domains in AMG proteins

Validating AMG protein sequences ensures that their identification on mVC genomes represents accurate annotations (i.e., predicted biological function). We used *in silico* approaches for protein validation by aligning AMG protein sequences with biochemically validated reference sequences from isolate bacteria or phages and assessed the presence or absence of functional domains and conserved amino acid residues. We highlighted cofactor coordination/active sites, cytochrome c motifs, substrate binding motifs, siroheme binding sites, cysteine motifs, and other

strictly conserved residues (collectively termed *residues*). Finally, we assessed if phage AMGs are under selection pressures to be retained.

Conserved residues identified on AMG protein sequences include: DsrA: substrate binding (R, KxKxK, R, HeR) and siroheme binding (CxgxxxC, CxxdC) (Supplementary Fig. 1); DsrC: strictly conserved cysteine motifs (CxxxgxpxpxxC) (Supplementary Fig. 2); SoxYZ: substrate binding cysteine (ggCs) and variable cysteine motif (CC) (Supplementary Fig. 3); SoxC: cofactor coordination/active sites (XxH, D, R, XxK) (Supplementary Fig. 4); SoxD: cytochrome c motifs (CxxCHG, CMxxC) (Supplementary Fig. 5). The identification of these residues on the majority of AMG protein sequences suggests they are as a whole functional. However, there are several instances of AMGs potentially encoding non-functional or distinctively different genes. For example, only 23 DsrC AMG protein sequences contained both of the strictly conserved cysteine motifs, 112 contained only the second cysteine motif, 1 contained only the first cysteine motif, and another 5 contained neither. The lack of strictly conserved cysteine motifs in phage DsrC has been hypothesized to represent AMGs with alternate functions during infection[140], but this hypothesis has yet to be validated. Likely, most DsrC AMG protein sequences lacking one or more cysteine residues functionally serve as TusE, a related sulfur transfer protein for tRNA thiol modifications[243]. Indeed, several mVCs originating from the human oral microbiome encode *tusE*-like AMGs that flank additional *tus* genes (Supplementary Fig. 2 and Supplementary Data 2). Further examples of missing residues include two mVCs encoding *soxD* in which one is missing the first cytochrome c motif, and both are missing the second cytochrome c motif (Supplementary Fig. 5). This initially suggests the presence of non-functional SoxD, but this notion is contested by the presence of conserved residues in SoxC. Functional SoxC, encoded adjacent to *soxD* in one of the mVCs, suggests that both likely retain function. It has been shown that phage proteins divergent

from respective bacterial homologs can retain their original anticipated activity or provide additional functions[244]. Overall, with the notable exception of 118 *tusE*-like AMGs, *in silico* analyses of AMG protein sequences suggests mVCs encode functional metabolic proteins.

To understand selective pressures on AMGs, we calculated the ratio of non-synonymous to synonymous nucleotide differences (*dN/dS*) in phage AMGs and their bacterial homologs to assess if phage genes are under purifying (stabilizing) selection. A calculated *dN/dS* ratio below 1 indicates a gene, or genome as a whole, is under selective pressures to remove deleterious mutations. Therefore, *dN/dS* calculation of mVC AMGs resulting in values below 1 would indicate that the viruses selectively retain the AMG's function by eliminating deleterious mutations in favor of those that provide function. Calculation of *dN/dS* for mVC *dsrA*, *dsrC* and *soxYZ* AMGs resulted in values below 1, suggesting AMGs are under purifying selection (Supplementary Fig. 6).

DSM AMGs likely manipulate key steps in sulfur oxidation pathways to redistribute energy

As previously stated, DSM AMGs encoded by the mVCs likely function specifically for the manipulation of sulfur transformations in the host cell during infection. To better understand the implications of this manipulation, we constructed conceptual diagrams of both sulfur (i.e., *dsr* AMGs) oxidation and thiosulfate (i.e., *sox* AMGs) oxidation/disproportionation in both uninfected and infected hosts (Fig. 2).

To understand the potential advantages of carrying *dsrC* and *dsrA* AMGs specifically, each step in the sulfide oxidation pathway needs consideration. During host-only sulfide oxidation[245], sulfide diffusing into the cell is converted into elemental sulfur by a sulfide:quinone oxidoreductase (e.g., *sqr*) and in some cases the pathway can begin directly with the import of elemental sulfur. The elemental sulfur can be stored in localized sulfur globules until it is

metabolized through the sulfide oxidation pathway[246]. During sulfide oxidation, elemental sulfur

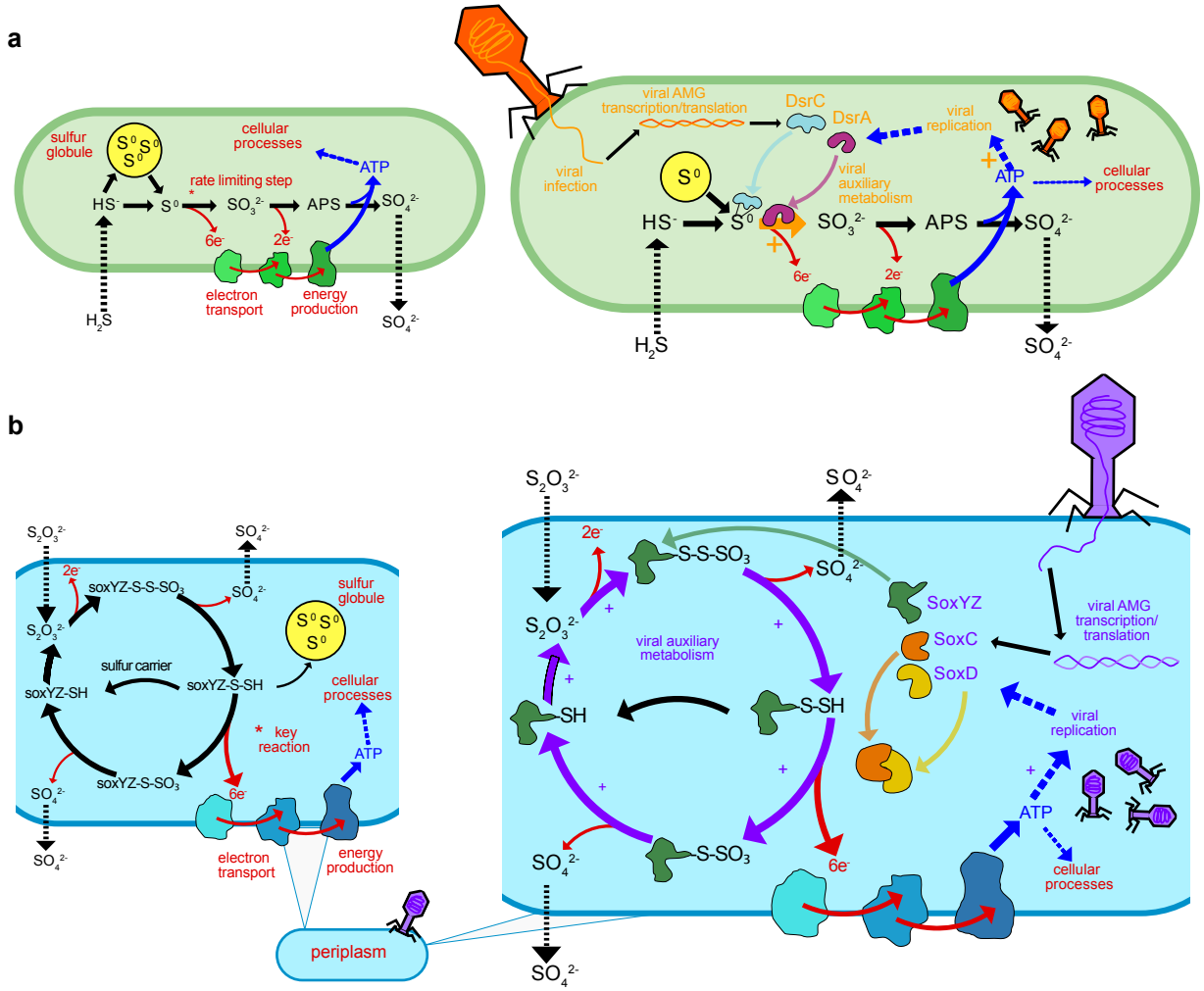


Fig. 2 Conceptual diagrams of viral DsrA, DsrC, SoxC, SoxD and SoxYZ auxiliary metabolism. **a** Microbial dissimilatory oxidation of hydrogen sulfide and stored inorganic sulfur. The resulting production of ATP utilized for cellular processes and growth and the pathway's rate limiting step is indicated with an asterisk (left). Viral infection and manipulation of sulfur oxidation by encoded DsrA or DsrC to augment the pathway's rate limiting step and increase energy yield towards viral replication (right). **b** Microbial dissimilatory oxidation of thiosulfate or storage of inorganic sulfur in the periplasm. The resulting production of ATP is utilized for cellular processes and the pathway's key energy yielding reaction indicated with an asterisk (left). Viral infection and manipulation of thiosulfate oxidation by encoded SoxC, SoxD or SoxYZ to augment the entire pathway and the key energy yielding step to increase energy yield towards viral replication (right). For **a** and **b** cellular processes are shown in red, sulfur oxidation pathway is shown in black, energy flow is shown in blue, and viral processes are shown in orange (**a**) or purple (**b**). For all pathway steps shown, microbial enzymes and sulfur carriers are functional in tandem with viral augmentation. APS, adenosine 5'-phosphosulfate.

carried by the sulfur carrier protein DsrC is oxidized into sulfite by the enzyme complex DsrAB. This step is estimated to be the rate limiting step in the complete pathway and yields the most electrons (six electrons) for ATP generation. Rate limitation is caused by either the saturation of the DsrAB enzyme complex or the DsrC carrier[247,248]. The final steps in sulfide/sulfur oxidation involve further oxidation of sulfite into adenosine 5-phosphosulfate (APS) and then sulfate by an APS reductase (e.g., *aprAB*) and sulfate adenylyltransferase, respectively (e.g., *sat*) which yields two electrons[245]. The obtained ATP can then be utilized for cellular processes. In contrast, during phage infection involving the modulation of sulfide oxidation, the rate limiting step (i.e., co-activity of DsrC and DsrA) can be supplemented by phage DsrC and/or DsrA to potentially increase the rate and ATP yield of the reaction as well as utilize any stored elemental sulfur[82]. This influx of ATP could then be effectively utilized for phage propagation (e.g., phage protein production, genome replication or genome encapsidation) (Fig. 2a).

Likewise, the normal state of thiosulfate oxidation/disproportionation may be augmented by phages encoding *soxYZ*, *soxC* and *soxD*. During host-only thiosulfate oxidation[249], thiosulfate is transported into the cell where the two thiol groups, transported by SoxYZ, undergo a series of oxidation reactions. A portion of the carried sulfur, after yielding two electrons, will be transported out of the cell as sulfate. The remaining carried sulfur may either be stored in elemental sulfur globules or proceed to the key energy yielding step. The key energy yielding step bypasses the storage of elemental sulfur and utilizes the SoxCD enzyme complex to produce six electrons for ATP yield[245,250]. During phage infection involving the modulation of thiosulfate oxidation/disproportionation, the entire pathway can be supported by both host and phage SoxYZ sulfur carriers in order to continuously drive elemental sulfur storage, which could then be oxidized by the Dsr complex. However, there is no evidence that phages benefit from coupling the *sox* and

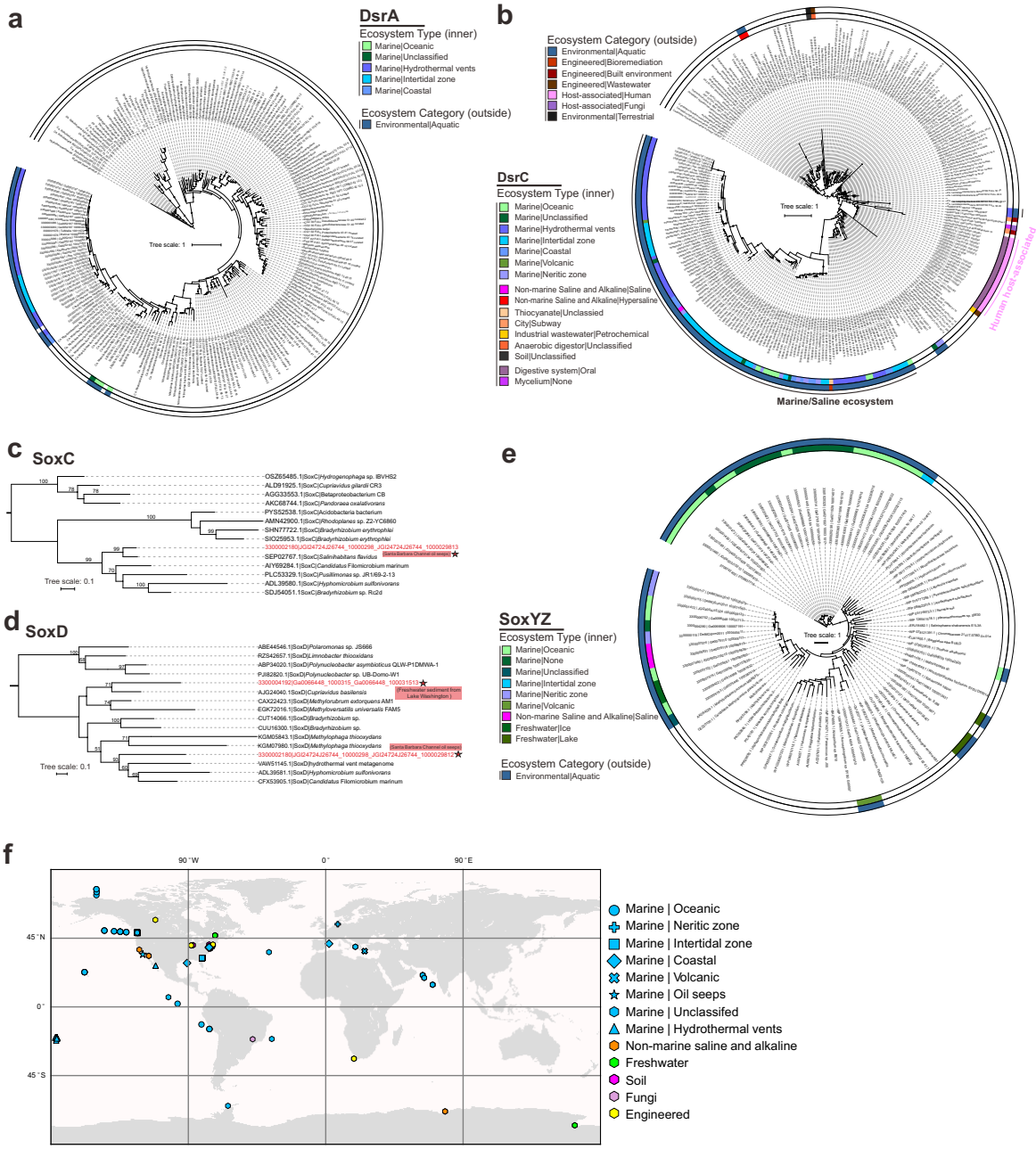


Fig. 3 Phylogenetic tree of AMG proteins and distribution of phage genomes (on a world map). **a**, **b** Phylogenetic trees of phage DsrA and DsrC **c**, **d**, **e** SoxC, SoxD, SoxYZ. Ultrafast bootstrap (UFBoot) support values (> 50%) are labelled on the nodes. **c**, **d** Phage gene encoded protein sequences are labeled with stars and their environmental origin information is labeled accordingly. The ecosystem type (inner ring) and ecosystem category (outside ring) are provided for phage genomes in the phylogenetic trees in **a**, **b**, **e**; different colors represent different ecosystem type and ecosystem category in the legends, blank places in each ring are for microbial reference. **f** World map showing distribution of phage genomes that contain the sulfur-related AMGs. Studies on human systems are excluded from the map. Different ecosystem types are represented by different symbols and colors in the legend.

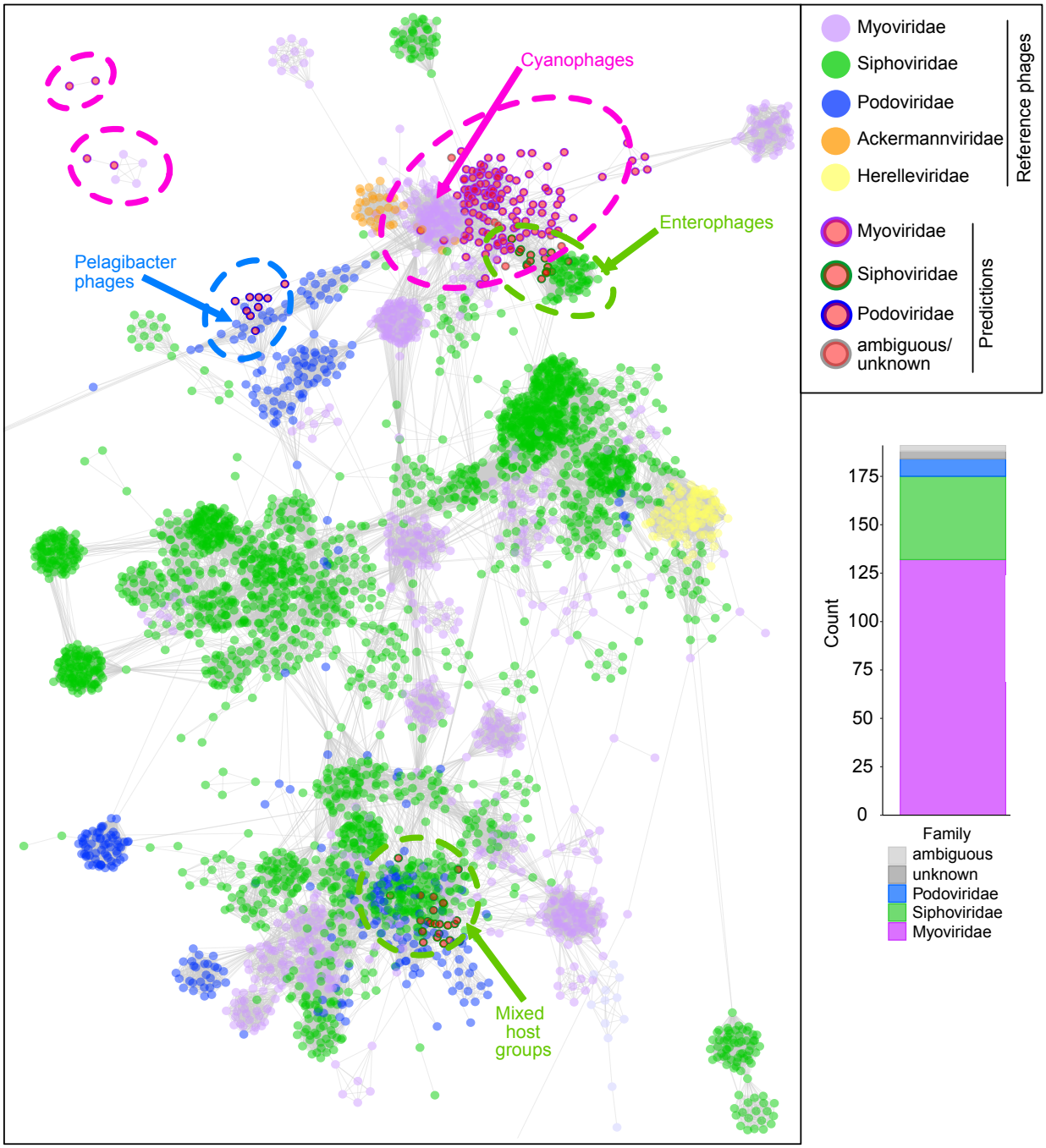


Fig. 4 Taxonomic assignment of mVCs and protein network clustering with reference phages. In the protein network each dot represents a single mVC (circles with outlines) or reference phage (circles without outlines), and dots are connected by lines respective to shared protein content. Genomes (i.e., dots) having more similarities will be visualized by closer proximity and more connections. Cluster annotations depicted by dotted lines were approximated manually. mVC taxonomy was colored according to predictions by a custom reference database and script, shown by bar chart insert.

dsr pathways since no mVCs were found to encode both a *sox* and *dsr* AMG simultaneously. Finally, phage SoxCD may be utilized to drive the pathway to the key energy yielding step. As with the *dsr* pathway, the resulting ATP would be utilized for phage propagation (Fig. 2b).

Sulfur phages are widely distributed in the environment

Next, we studied the ecological and distribution patterns of mVCs encoding DSM AMGs. We characterized their diverse ecology and distribution patterns in various environments by building phylogenetic trees using the identified AMG and reference microbial proteins, and parsing environmental information of mVC metadata from the IMG/VR database. We identified mVCs encoding *dsrA* mainly in a few ocean environments, while more widely distributed mVCs encoding *dsrC* were found in in ocean, saline, oil seep-associated, terrestrial, engineered, and symbiotic environments (Fig. 3a, b). For *soxC* and *soxD*, we only identified mVCs encoding these AMGs in two metagenome datasets, one from Santa Barbara Channel oil seeps (mVC encoding both *soxC* and *soxD*) and another from freshwater sediment from Lake Washington (Fig. 3c, d). The mVCs encoding *soxYZ* were discovered in aquatic environments, consisting of different ocean, saline and freshwater ecosystem types (Fig. 3e). In addition to mVC distribution amongst diverse ecosystem types we identified wide biogeographic distribution across the globe (Fig. 3f). Collectively, these DSM AMGs are ecologically and biogeographically ubiquitous, and potentially assist host functions in many different environment types and nutrient conditions (including both natural and engineered environments).

Sulfur phages are taxonomically diverse within the order Caudovirales

We applied two approaches to taxonomically classify and cluster the identified mVCs. First, we used a reference database similarity search to assign each mVC to one of 25 different prokaryote-infecting viral families (see Methods). The majority of mVCs were assigned to *Myoviridae* (132 mVCs; 69%), *Siphoviridae* (43 mVCs; 22%) and *Podoviridae* (9 mVCs; 5%). These three families represent dsDNA phages belonging to the order *Caudovirales*. The remaining seven mVCs were identified as ambiguous *Caudovirales* (3 mVCs; 1.5%) and unknown at both the order and family levels (4 mVCs; 2%). However, based on the data presented here and previous classifications[35,82,140], the seven unclassified mVCs likely belong to one of the three major *Caudovirales* families (Fig. 4).

In accordance with these results we constructed a protein sharing network of the mVCs with reference viruses from the NCBI GenBank database (Fig. 4). The mVCs arranged into four main clusters with reference *Myoviridae*, *Siphoviridae* and *Podoviridae*, and four individual mVCs were arranged outside of main clusters. Of the seven mVCs with ambiguous/unknown predictions, six clustered with *Myoviridae* and *Siphoviridae* mVCs and reference phages, further suggesting their affiliation with major *Caudovirales* families. Overall, the network diagram validated the reference-based taxonomic assignment results (i.e., mVCs predicted to be podoviruses clustered with reference podoviruses, as with myoviruses and siphoviruses). On the basis of these findings, we hypothesize that the function(s) of DSM AMGs during infection is most likely constrained by specific host sulfur metabolisms rather than viral taxonomy. The broad distribution of DSM AMGs across *Caudovirales* further suggests that this modulatory mechanism is established across multiple taxonomic clades of phages, either arising independently or acquired via gene transfer. Most mVCs clustered with reference phage genomes of varying taxonomy and host ranges, though there was not significant enough protein similarity between the mVCs and these reference phages

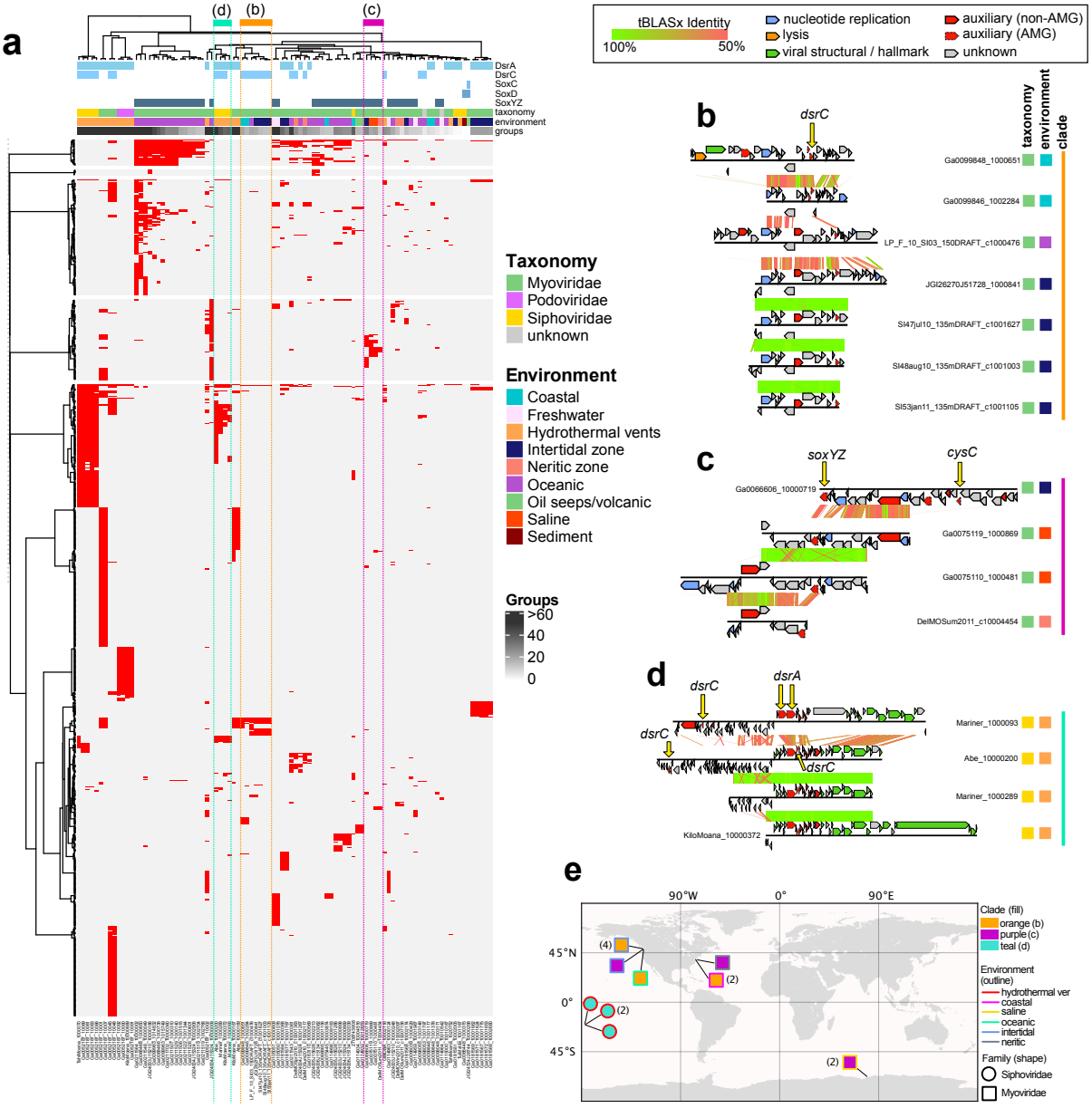


Fig. 5 mVC protein grouping and genome alignments. **a** mVC hierarchical protein grouping where each row represents a single protein group (887 total) and each column represents a single mVC (94 total). Metadata for encoded AMGs, estimated taxonomy, source environment and number of protein groups per mVC is shown. Clades respective of **b**, **c** and **d** are depicted by colored dotted lines. Genome alignments of **b** seven divergent Myoviridae mVCs encoding *dsrC* from diverse environments, **c** four divergent Myoviridae mVCs encoding *soxYZ* from diverse environments, and **d** four divergent Siphoviridae mVCs encoding *dsrA* and *dsrC* from hydrothermal environments. For the genome alignments, each black line represents a single genome and arrows represent predicted proteins which are colored according to VIBRANT annotations; genomes are connected by lines representing tBLASTx similarity. **e** Map of geographic distribution of 15 mVCs depicted in **b**, **c** and **d**, annotated with respective clade, source environment and taxonomic family.

to suggest similarity at the genus level. It is likely that host range stems beyond these indicated taxa, suggested by the inclusion of a SUP05-infecting mVC[82] within the *Pelagibacter* cluster. In the present state of the reference databases, this type of protein sharing network cannot be used to reliably predict the host range of these uncultivated mVCs, but rather is indicative of shared taxonomy (e.g., mVC podoviruses clustering with reference podoviruses, and likewise for myoviruses and siphoviruses). Based on phylogeny and AMG protein similarity, the mVC host range appears to be primarily Gammaproteobacteria from the SUP05/*Thioglobus* clades, with the possibility of extended host range to Methylophilaceae in the Betaproteobacteria (Fig. 3). Using CRISPR analysis against 7,178 spacers from 25 metagenomes we were unable to validate any mVC link to a putative host.

Sulfur phages display diversification across environments and genetic mosaicism

To further assess the diversity of the identified mVCs and their evolutionary history, we analyzed shared protein groups as well as gene arrangements between individual mVCs. All predicted proteins from 94 of the mVCs, excluding mVCs encoding only *tusE*-like AMGs, were clustered into protein groups. Our protein clustering method for featuring the diversity of the mVCs, despite representing partial genome sequences, was assessed and verified using Caudovirales phages from NCBI RefSeq (see Methods). A total of 794 protein groups representing 3677 proteins were generated, roughly corresponding to individual protein families. Only a few protein groups were globally shared amongst the mVCs, including common phage proteins (e.g., *phoH*, *nifU*, *iscA*, nucleases, helicases, lysins, RNA/DNA polymerase subunits, ssDNA binding proteins and morphology-specific structural proteins) (Fig. 5a). A lack of shared protein groups between the mVCs may be anticipated due to missing genes on the partial mVC scaffolds.

However, distinct phage lineages share few protein groups regardless of genome completeness. Overall, the results of the protein grouping are consistent with that of taxonomic clustering, further highlighting the diversity of phage genomes that encode DSM AMGs. A lack of universally shared protein groups likewise suggests the DSM AMGs function independently of other host metabolic pathways and likely strictly serve to supplement host DSM pathways.

To identify if the shared protein groups are relevant to the DSM AMGs and further highlight mVC diversity we generated a second set of protein clusters corresponding to five proteins before and five proteins after the DSM AMG, including the AMG. Since the true completeness of the entire mVC cannot be determined, this subset of 11 proteins adjacent to the DSM AMGs was utilized to best represent potential shared features regardless of completion. This second set included 70 mVCs (we excluded 24 mVCs for which the encoded DSM AMG was within five genes of a scaffold end). In total, 116 protein clusters were generated (Supplemental Figure 7). Interestingly, nearly identical proteins were common, namely PhoH, NifU, IscA, GrxD, TusA, NrdAB, RNA/DNA polymerase subunits, ssDNA binding proteins and morphology-specific structural proteins. However, the shared groups represented only a small subset of all groups. Therefore, in addition to the common functions such as iron-sulfur cluster formation (e.g., NifU, IscA, GrxD and TusA), the mVCs encode dissimilar proteins, likely resulting from varied evolutionary backgrounds.

Most mVCs that formed clades according to whole mVC shared protein groups could be explained by shared taxonomy and/or source environment. This observation further validates that despite the mVCs representing partial scaffolds, they encoded sufficient information to be accurately grouped into clades. That is, similar mVCs by genome alignment, as with taxonomy and source environment, were found to group into the same clade. For example, 16 Myoviridae

mVCs encoding *soxYZ* from oceanic environments clustered together, only differing according to their total number of representative protein groups (Fig. 5a). There were exceptions, such as seven *dsrC*-encoding mVCs which displayed variable pairwise protein similarity (at a 50% identity cutoff) and variation in the location of their *dsrC* gene within their genome, despite a clearly shared and distinctive synteny of other genes (Fig. 5b). The seven mVCs originated from three different marine environment types (coastal, oceanic and intertidal) and were all predicted to be myoviruses (Fig. 5b). This diversity is likely explained by the retention of the *dsrC* gene over time despite components of the genome undergoing genetic exchange, recombination events or mutation accumulation. Phages are well known to display genetic mosaicism, or the exchange and diversification of genes and gene regions[242,251]. The same conclusion can be made with myoviruses encoding *soxYZ* from different marine environments (intertidal, saline and neritic) (Fig. 5c) as well as siphoviruses encoding both *dsrC* and *dsrA* from hydrothermal environments (Fig. 5d). In addition to distribution amongst diverse environmental categories these genetically mosaic mVCs, per protein sharing clade, are geographically dispersed (Fig. 5e). Additionally, one mVC (Ga0066606_10000719) encoding *soxYZ* also encodes the assimilatory sulfur metabolism AMG *cysC* (Fig. 5b). This presents an interesting discontinuity suggesting that this particular mVC, as well as three others encoding *cysC* (Ga0052187_10001, Ga0052187_10007 and JGI24004J15324_10000009), target both dissimilatory and assimilatory sulfur metabolism simultaneously to more generally affect sulfur metabolism in the host.

Estimations of sulfur phage contributions to sulfur oxidation based on omics-data analyses

We utilized metagenomic datasets containing the mVCs to calculate the ratio of phage:total genes for each AMG. The phage:total gene ratios within a community and for each predicted

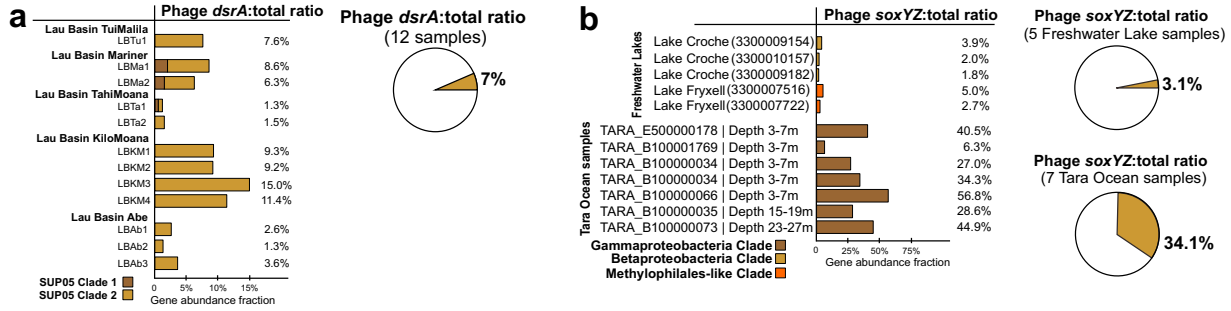


Fig. 6 Phage to total *dsrA* and *soxYZ* gene coverage ratios. **a** Phage *dsrA* to total (phage and bacterial *dsrA* gene together) gene coverage ratios. The contribution of phage *dsrA* genes from different SUP05 Gammaproteobacteria clades is shown in different colors. The average phage *dsrA*:total ratio was calculated from 12 samples. **b** Phage *soxYZ* to total gene coverage ratios. The contribution of phage *soxYZ* genes from three different clades is shown in different colors. Genes from freshwater lake and *Tara* Ocean samples were compared separately, and the average phage *soxYZ*:total ratios were calculated and compared separately as for freshwater lake and *Tara* Ocean samples. *Tara* Ocean sample IDs were labeled and the corresponding metagenome IDs were listed in Supplementary Data 4. LBTu, Lau Basin Tui Malila; LBMa, Lau Basin Mariner; LBTa, Lau Basin Tahi Moana; LBKM, Lau Basin Kilo Moana; LBAb, Lau Basin Abe.

phage-host pair can be used to estimate phage contributions to sulfur and thiosulfate oxidation/disproportionation. This relies on the assumption that gene ratios can proportionally reflect real metabolic activities, and that host cells in a virocell state retain the same level of environmental fitness as compared to uninfected microorganisms. By mapping metagenomic reads to AMGs and putative bacterial hosts within the metagenome, we obtained the mVC AMG to total gene ratios, which represents the relative contribution of AMG functions to the representative metabolism such as sulfur oxidation (Supplementary Data 3, 4, Supplementary Fig. 8). We calculated mVC *dsrA* (Fig. 6a) and *soxYZ* (Fig. 6b) gene coverage ratios in hydrothermal, freshwater lake, and *Tara* Ocean metagenomic datasets. We identified phage-host gene pairs which contained mVC AMGs and their corresponding host genes from the phylogenetic tree of DsrA and SoxYZ (Supplementary Figs. 9, 10). Our results show that phage *dsrA* contributions in hydrothermal environments arise primarily from the SUP05 Gammaproteobacteria Clade 2; and those of phage *soxYZ* are niche-specific, with Lake Croche, Lake Fryxell, and *Tara* Ocean samples

mainly represented by the Betaproteobacteria Clade, Methylophilales-like Clade, and Gammaproteobacteria Clade, respectively. This indicates the specificity of AMGs being distributed and potentially functioning in each environment. The average phage:total gene coverage ratios also differ in individual groups, with phage *soxYZ*:total ratio in *Tara* Ocean samples being the highest (34%), followed by phage *dsrA*:total ratio in hydrothermal samples (7%) and phage *soxYZ*:total ratio in freshwater lakes (3%). Phage *soxYZ*, the sulfur carrier gene, in the oceans have higher phage:total gene coverage ratio compared to *dsrA*, a component of the catalytic core of Dsr complex, in the other two environments. *Tara* Ocean samples used here are all from epipelagic zones characterized as oxygenated layers with low concentrations of sulfur[252], while plume samples (Lau Basin) are from deep-ocean hydrothermal ecosystems with high concentrations of sulfur[253]. Nevertheless, along with observations associated with phage *dsrC*, our results suggest that AMGs encoding sulfur carriers rather than catalytic subunits appear to be more favored by phages. These findings were unexpected since we expected epipelagic environments to have lower sulfur oxidation activity in comparison to hydrothermal plumes. While the limited environment types, conditions and sulfur AMGs studied here do not provide sufficient statistical confidence to generalize these results, especially when comparing different genes from separate environments, higher abundance of sulfur carrier genes in phage nevertheless could still be a common phenomenon. Additionally, although gene abundance ratios do not necessarily represent function contributions, this scenario still provides a reasonable estimation to suggest considerable sulfur-oxidizing contributions of phage sulfur AMGs in virocells.

Subsequently, the phage:host AMG coverage ratios for individual phage-host pairs were calculated to estimate the potential functional contribution within each environmental sample (Figs. 7a, b, Supplementary Data 3, 4, Supplementary Figs. 11, 12). By taking average ratios of

groups of *dsrA* phage-host pairs in SUP05 Clade 1 and SUP05 Clade 2, and *soxYZ* phage-host pair in freshwater lake and *Tara* Ocean samples, we found that within each pair the phage:total gene coverage ratios were generally higher than ~50%. These within-pair phage:total gene coverage ratios are much higher than the above phage:total ratios in the whole community. *Tara* Ocean samples also have the highest average phage:total gene coverage ratios of phage-host pairs among these three environments, as with the pattern of ratios in the whole community. To estimate the percentage of virocells in the community, we use average values of 16% and 15% for marine (range of 3-31% or 3-26% in free-living and particulate-associated marine bacteria)[37,254] and freshwater lake (range of 1 to 17%+/-12%)[255] bacteria . The estimated phage:total gene coverage ratio within the whole community should be the virocell percentage multiplied by the average phage:total gene coverage ratio within phage-host pairs (as the phage gene coverage), and then divided by total gene coverage. We found that estimated ratios are not consistent with the observed ratios (*Tara* Ocean: estimated ratio 68% versus the observed ratio of 34%; hydrothermal environment: estimated ratio of 15-38% versus observed ratio of 12-20%; freshwater lake: estimated ratio of 27% versus observed ratio of 3.1%). This could result from an unknown fraction of the host cells being infected by phages that do not contain DSM AMGs as these virocells do not contribute phage genes to sulfur metabolism and/or the percentages of cells in a virocell state being below the average levels.

The above analyses suggest that DSM AMGs likely contribute significantly to function of host-driven metabolisms on the scale of both community level and individual phage-host pairs, while the ratio of contribution varies greatly for each environment and each niche-specific AMG. Importantly, phage-encoded *soxYZ* have a high gene coverage contribution to pelagic ocean microbial communities, which highlights the functional significance of phage-driven sulfur

cycling metabolisms, and that of thiosulfate oxidation/disproportionation as a whole in this environment, which remains critically under-studied[62,140].

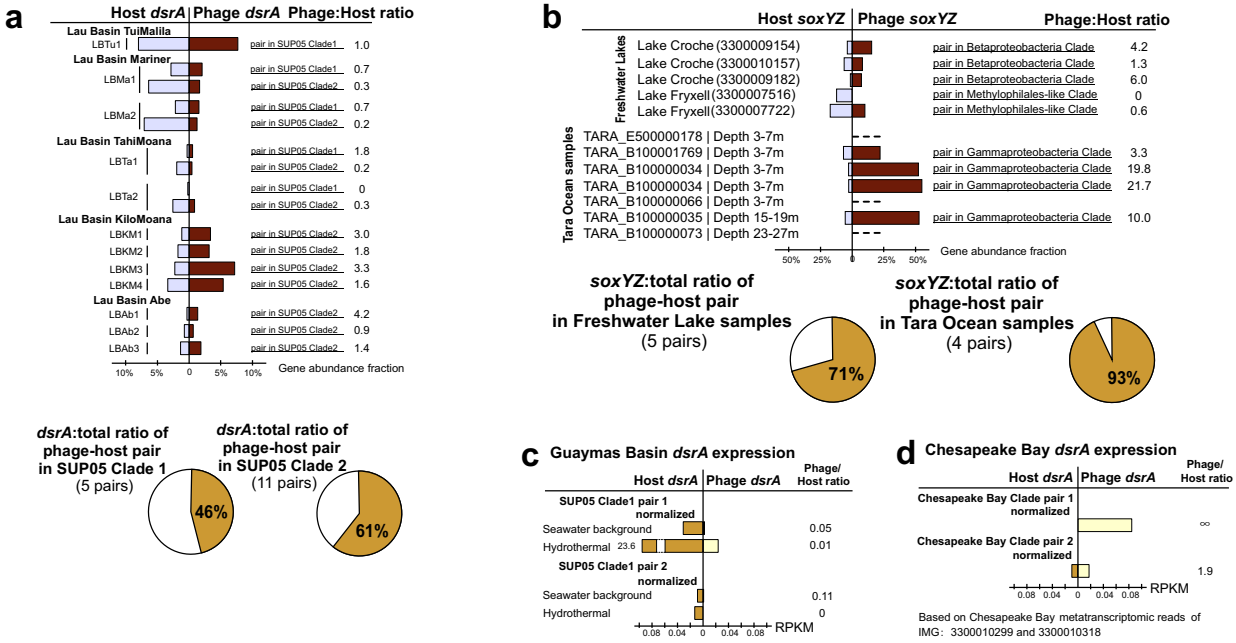


Fig. 7 Phage to host *dsrA* and *SoxYZ* gene coverage ratios and *dsrA* gene expression comparison between phage and host pairs. **a** Phage *dsrA* to total gene coverage ratios of each phage-host pair. Average phage *dsrA*:total ratios of phage-host pairs in SUP05 Clade 1 and Clade 2 were calculated by 5 and 11 pairs of genes, respectively. **b** Phage *soxYZ* to total gene coverage ratios of each phage-host pair. The contribution of phage *soxYZ* genes from three different clades is shown in different colors. Average phage *dsrA*:total ratios of phage-host pairs in freshwater lakes and *Tara* Ocean were calculated separately. *Tara* Ocean sample IDs were labeled and the corresponding metagenome IDs were listed in Supplementary Data 4. **c** Phage to host *dsrA* gene expression comparison in Guaymas Basin metatranscriptomes. The same database was used for mapping both hydrothermal and background metatranscriptomic datasets **d** Phage to host *dsrA* gene expression comparison in Chesapeake Bay metatranscriptomes. The same database was used for mapping all Chesapeake Bay metatranscriptomic datasets. Gene expression levels are shown in RPKM normalized by gene sequence depth and gene length. LBTu, Lau Basin Tui Malila; LBMa, Lau Basin Mariner; LBTa, Lau Basin Tahi Moana; LBKM, Lau Basin Kilo Moana; LBAb, Lau Basin Abe; RPKM, reads per kilobase per million mapped reads.

Rapid alteration of sulfur phage dsrA activity across geochemical gradients

Since DSM AMGs are associated with critical energy generating metabolism in microorganisms, we wanted to study the ability of sulfur phages to respond to changing geochemistry, involving virocell-driven biogeochemical cycling. In hydrothermal ecosystems,

reduced chemical substrates such as H_2S, S^0, CH_4, and H_2 display sharp chemical gradients as they are released from high-temperature vents and dilute rapidly upon mixing with cold seawater. Microorganisms in deep-sea environments respond to such elevated concentrations of reduced sulfur compounds by upregulating their metabolic activity in hydrothermal environments[253,256]. These characteristics make hydrothermal and background deep-sea environments a contrasting pair of ecological niches to investigate alteration of AMG expression. We used transcriptomic profiling to study gene expression in phage:host pairs recovered from hydrothermal vents in Guaymas Basin and background deep-sea samples in the Gulf of California (Supplementary Data 3, Supplementary Figs. 11, 12). Sulfur phage *dsrA* expression measured in reads per kilobase of transcript (RPKM) varied from 0.03-3 in the background deep-sea to 0.40-39 in hydrothermal environments (Supplementary Data 3). Average phage *dsrA* expression ratio of hydrothermal to background was 15 (Supplementary Data 3). Limited by coding gene repertoire and their biology, phages themselves do not have the ability to independently sense and react to sulfur compounds. However, our results suggest that sulfur phage activities, occurring within a virocell, are closely coupled to changing geochemistry with higher observed activity in environments with greater concentration of reduced sulfur compounds.

In Guaymas Basin hydrothermal environments, as reflected by two pairs of SUP05 Clade 1 phage and host *dsrA* genes, phage to host *dsrA* transcript ratios varied from 0 to 0.11 (Fig. 7c). In contrast, in Chesapeake Bay, as reflected by two pairs of phage and host *dsrA* transcripts (Chesapeake Bay *dsrA* clade), phage to host *dsrA* transcript ratios varied from 1.9 to infinity (host transcript abundance is zero). The low abundance of phage *dsrA* in hydrothermal metatranscriptomes is in sharp contrast to the high abundance of phage *dsrA* in hydrothermal metagenomes (observed at Guaymas Basin and Lau Basin) (Fig. 7a, c). One explanation for this

observation is that this scenario could be an accident but not representative of real phage gene expression patterns in hydrothermal systems, possibly occurring in a situation when phage activity was very high just prior to sampling. In this scenario, the majority of hosts/virocells might have lysed post viral infection.

Discussion

Since the first descriptions of viral metabolic reprogramming using AMGs[73] there has been interest in the extent and overall impact of viral auxiliary metabolism on global energy flows and ecosystem nutrient availability[257]. Through metagenomic surveys and investigation, we have expanded the current understanding of viral auxiliary metabolism impacting dissimilatory sulfur oxidation processes. Specifically, we have shown that diverse lineages of phages are involved in these processes, investigated their biogeography, ecology, and evolutionary history, and estimated their potential effects on microbiomes. From this, several hypotheses and questions regarding viral auxiliary metabolism and sulfur cycling can be addressed.

First, our findings support previous hypotheses that viral metabolism targets key or bottleneck steps in host metabolic pathways. DsrA, DsrC, SoxYZ, SoxC, and SoxD all alleviate bottlenecks in sulfur and thiosulfate oxidation/disproportionation[82,258]. We did not identify other genes in sulfur oxidation pathways such as sulfide:quinone oxidoreductase, flavocytochrome *c* cytochrome/flavoprotein subunits, APS reductase subunits, sulfate adenylyltransferase, *dsrB*, or *soxAB* for other necessary steps of sulfur oxidation. However, this poses the additional question of why DsrB, the dimer pair to DsrA, has yet to be identified as an AMG. Furthermore, sulfur carriers, rather than enzymes, appear to be more favored by phages. In total, 174 mVCs in this study encoded at least one sulfur carrier (*dsrC*, *tusE*-like, *soxYZ*) with only the remaining 17

encoding catalytic subunits of enzymes (*dsrA*, *soxC*, *soxD*). Phage sulfur carriers were observed to be more abundant in the phage community than catalytic subunits such as *dsrA*. This may be due to the greater need for sulfur carriers (e.g., *dsrC*) to drive dissimilatory sulfur transformations. Evidence for this hypothesis is provided by observations that sulfur carriers are often constitutively expressed in host cells in comparison to respective catalytic components (e.g., *dsrA*)[248,259]. By providing transcripts and proteins of these important pathway components during infection, phages encoding DSM AMGs may benefit more from obtaining greater energy and self-catalyzing substrates within a virocell.

The data presented by mVC protein clustering and genome alignments (Fig. 5) supports the hypothesis that the DSM AMGs are retained on fast evolving phage genomes, pointing specifically to a role of the AMG in increasing phage replication abilities and fitness. Although the mechanism of dispersion is unknown for most of the mVCs it is likely that a single AMG transfer event occurred within each clade based on retention of similar gene arrangements at AMG locations in the respective genomes. This suggests that the AMG were retained despite niche (i.e., geographic and environmental) differentiation of individual mVC populations. It has been postulated that AMGs, like other phage genes, must provide a significant fitness advantage in order to be retained over time on an evolving phage genome[74].

Taken together, these observations support the conclusion that viral auxiliary metabolism targets key steps in host metabolic pathways for finely tuned, host-dependent manipulation of energy production or nutrient acquisition. Although the fitness effects of DSM AMGs have not been quantified in a model system, the geographical distribution of identified mVCs and retention of AMGs by phages despite constrained coding capacity strongly suggests a significant fitness benefit of encoding DSM AMGs. The exact fitness benefit achieved from encoding DSM AMGs

remains elusive without cultured representatives of phage-host pairs and subsequent genetic manipulation abilities. Furthermore, a model system would be beneficial for elucidating the functionality of the AMGs, beyond the evidence from protein domain analyses presented here. Although most AMGs encoded conserved functional domains and residues the identification of divergent sequences, such as *tusE*-like AMGs encoding a single cysteine residue or *soxD* AMGs that appear to lack a cytochrome c motif, necessitates further biochemical evaluation of AMG-encoded proteins. For example, a divergent PebA encoded by a cyanophage was found to short-circuit the original host pathway by excluding the necessity of the subsequent host enzyme PebB[244]. It is possible sulfur AMG-encoded proteins likewise short-circuit or increase the rate of host sulfur oxidation pathways using divergent AMGs.

Since DSM AMGs have been identified on phages from all three major *Caudovirales* families it is likely that the fitness benefits deal specifically with sulfur oxidation and electron yield from bolstering the speed or efficiency of the pathway, rather than phage taxonomy-dependent reasons. Based on evidence from systems with cyanophages encoding photosystem AMGs, a potential utility of sulfur AMGs in bolstering the speed or efficiency of the pathway would be to increase the yield rate of dNTPs for genome replication[75]. Further evidence would suggest the AMGs could also function to upregulate the expression of important metabolic genes that encode for unstable protein products[260]. In such cases the host cell is adapted for such metabolic constraints but the replication rate of the phage is directly dependent on the translation rate of the given AMG protein product[260]. Therefore, the phage, rather than the host, benefits from an additional copy of the metabolic gene leading to recombination and retention of the AMG on the phage genome. It is most likely that the phages benefit primarily in the short term and during active lytic infection due to the abundance of DSM AMGs on lytic phage genomes. Yet, the

presence of assimilatory sulfate reduction genes (i.e., *cysC*) in conjunction with DSM genes provides an example of a possible exception with a more general sulfur manipulation, highlighting the necessity of further investigations into viral auxiliary metabolism.

The abundance of phage DSM AMGs in metagenomes and metatranscriptomes as measured by phage:total gene coverage ratios suggest that phage-mediated reduced sulfur transformations can contribute significantly to fluxes and budgets of sulfur within the community (Fig. 8, Supplementary Figs. 8, 13). Within each phage-host pair, phage genes contribute to over half of gene coverage associated with the sulfur and thiosulfate oxidation pathways, which highlights the underappreciated role of phages encoding DSM AMGs in remodeling sulfur cycling, especially for the oxidation of reduced sulfur. Reduced sulfur compounds such as H_2S, S^0, and $S_2O_3^{2-}$ are abundant in hydrothermal systems with hydrothermal fluids at Guaymas Basin containing aqueous H_2S concentrations of up to ~6 mmol/kg (endmember measurement), while that of background seawater is negligible[256,261]. Previously reported estimates of energy budgets for sulfur oxidizing bacteria in the Guaymas Basin hydrothermal system suggest that up to ~3900 J/kg is available for microbial metabolism, of which up to 83% may derive from sulfur oxidation[256]. Sulfur phage *dsrA* expression levels (arising from virocells) were elevated in hydrothermal systems in comparison to the background deep-sea, hinting at significant contributions of virocells mediating phage-driven sulfur oxidation to the overall energy budget. Assuming that in Guaymas Basin the phage:total *dsrA* gene coverage ratio is 10% (the average level in Lau Basin hydrothermal environments), it may be estimated that ~320 J/kg of energy for microbial metabolism from hydrothermal vent fluids may in fact be transformed by sulfur AMGs. Although the majority of host manipulation and lysis by phages likely occurs in the absence of AMGs, we show that phages encoding sulfur AMGs can be a direct component of the

sulfur biogeochemical cycle with the ability to manipulate microbial metabolism associated with multiple reduced sulfur compounds. This direct manipulation may impact sulfur budgets at ecosystem scales. It is therefore essential that future assessments of biogeochemical cycling incorporate the role of phages and their impacts on sulfur pools. Limited by the resolution of omics-based approach in this study, finer scale phage-host interactions and activities could not be achieved, which justifies the necessity to

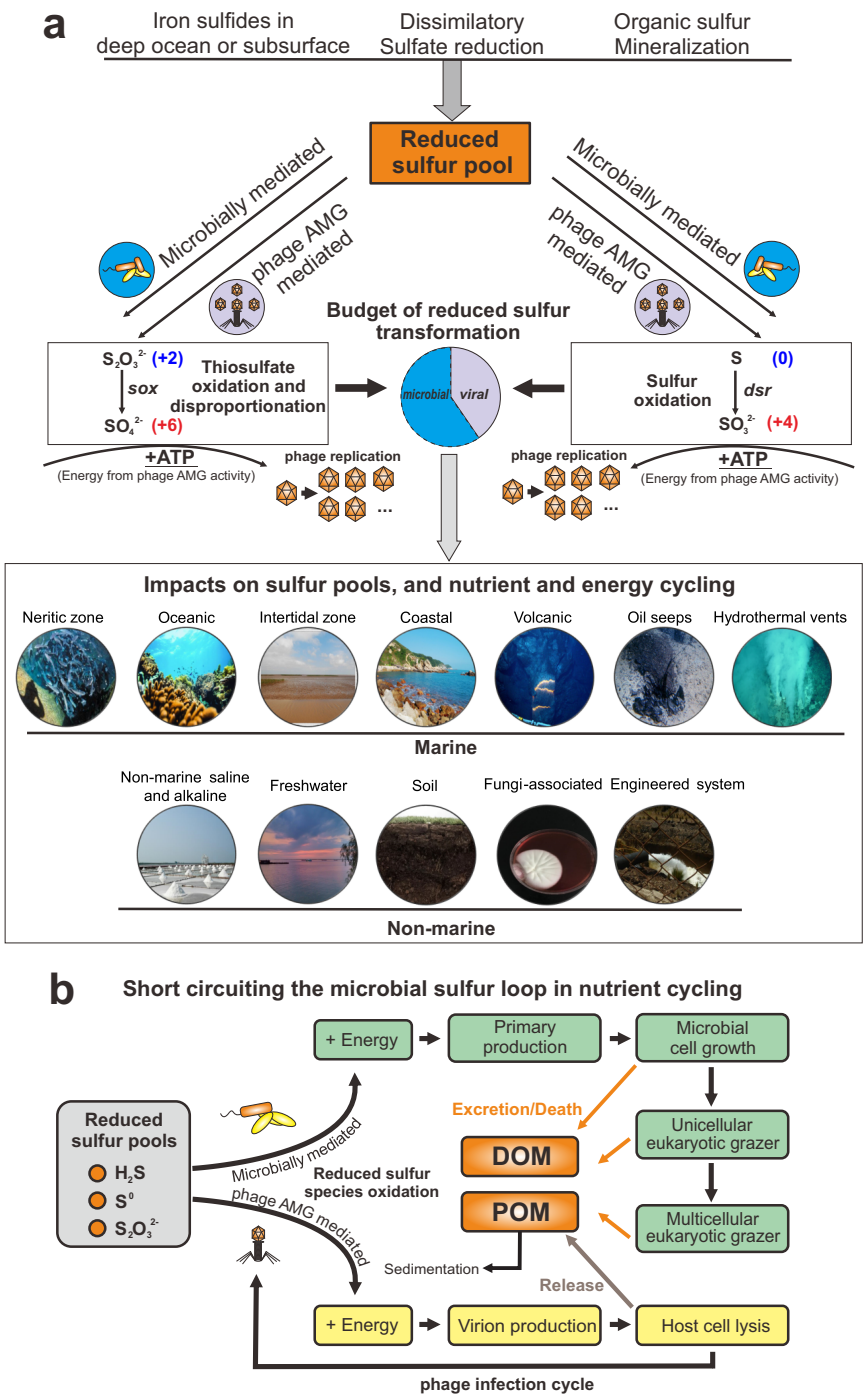


Fig. 8 Conceptual figure indicating the ecology and function of AMGs in sulfur metabolisms. **a** DSM AMG effect on the budget of reduced sulfur transformation. **b** Diagram of virus-mediated metabolism short circuiting the microbial sulfur loop in nutrient cycling. DOM, dissolved organic matter; POM, particulate organic matter.

reinforce fine-scale phage AMG activity research within host cells in future.

Across diverse environments on the Earth, the reduced sulfur pool includes sources of deep ocean or subsurface deposited iron sulfides, and reduced sulfur species from dissimilatory sulfate reduction and organic sulfur mineralization (Fig. 8a). Sulfur phage AMG-assisted metabolism contributes to the redistribution of sulfur-generated energy and can alter its budgets, which have so far only been attributed to microbial processes (Fig. 8a). Within virocells, phage mediated sulfur oxidation will take advantage of gene components of sulfur-metabolizing pathways, express transcripts, and produce enzymes to re-direct energy for the use of phage replication (Fig. 8a). Globally distributed sulfur phages are widely distributed across various environments and impose significant impacts on the sulfur pools, as well as nutrient and energy cycling (Fig. 8a). At the same time, phage AMG mediated sulfur oxidation can short-circuit the microbial sulfur loop from reduced sulfur pools to dissolved and particulate organic matter (DOM/POM) (Fig. 8b). Without viral infection, energy generated by reduced sulfur pools would typically be used for primary production to fuel microbial cell growth, and then transferred higher up the food chain to grazers. Through cell excretion effects, cell death and nutrient release, DOM/POM produced from sulfur-based primary production would be released to the environment. However, during infection by sulfur phages, energy generated in virocells by reduced sulfur pools could be used towards phage reproduction and propagation. After virion production and packaging, lytic phages would lyse the host cell, and release DOMs into the environment. This DSM AMG mediated approach thereby short-circuits the microbial sulfur loop. Additionally, POM generated by reduced sulfur oxidizing processes could also be sequestered into the carbon pool deposited in the deep subsurface. It is not clear how and to what extent phage would change carbon cycling landscape between sequestrated carbon and bioavailable carbon, while it is certain that the change caused by phage AMG

metabolism should be explicitly addressed in the future in the context of global biogeochemical cycle and climate change.

In conclusion, we have described the distribution, diversity and ecology of phage auxiliary metabolism associated with sulfur and demonstrated the abundance and activity of sulfur phages in the environment. Yet, many questions remain unanswered. Future research will involve unraveling mechanisms of sulfur phage and host interaction, remodeling of sulfur metabolism at the scale of individual virocells, microbial communities and ecosystems, and constraining sulfur budgets impacted by sulfur phages.

Methods

mVC acquisition and validation

The Integrated Microbial Genomes and Virome (IMG/VR) database[141,262] (v2.1, October 2018) was queried for dissimilatory sulfur metabolism genes using trusted cutoffs of custom built HMM profiles[132]. A total of 192 unique mVCs greater than 5kb in length were identified that encoded *dsr* or *sox* gene(s). For consistency between these mVCs, open reading frames were predicted using Prodigal (-p meta, v2.6.3)[98]. Each of the 192 mVCs were validated as phage using VIBRANT[117] (v1.2.1, virome mode), VirSorter[115] (v1.0.3, virome decontamination mode, virome database) and manual validation of viral hallmark annotations (Supplementary Data 5). To identify lysogenic mVCs, annotations were queried for the key terms “integrase”, “recombination”, “repressor” and “prophage”. Annotations of validated mVCs are provided in Supplementary Data 2. Five mVCs not identified by either program were manually verified as phage according to VIBRANT annotations (i.e., KEGG, Pfam and VOG databases) by searching for viral hallmark genes, greater ratio of VOG to KEGG annotations and a high proportion of unannotated proteins.

Note, not all 192 mVCs were predicted as phage by VIBRANT, but all mVCs were given full annotation profiles. One scaffold was determined to be non-viral and remove based on the presence of many bacterial-like annotations and few viral-like annotations. Validation (including software-guided and manually inspected procedures) produced a total of 191 mVCs encoding 227 DSM AMGs. It is of note that the DSM AMGs carried by three mVCs (Ga0121608_100029, Draft_10000217 and Ga0070741_10000875) could not be definitely ruled out as encoded within microbial contamination. This was determined based on the high density of non-phage annotations surrounding the AMGs in conjunction with the presence of an integrase annotation, suggesting the possibility of phage integration near the AMG.

Taxonomy of mVCs

Taxonomic assignment of mVCs was conducted using a custom reference database and script. To construct the reference database, NCBI GenBank[210] and RefSeq[207] (release July 2019) were queried for "prokaryotic virus". A total of 15,238 sequences greater than 3kb were acquired. Sequences were dereplicated using Mash[263] (v2.0) and Nucmer[264] (v3.1) at 95% sequence identity and 90% coverage. Dereplication resulted in 7,575 sequences. Open reading frames were predicted using Prodigal (-p meta, v2.6.3) for a total of 458,172 proteins. Taxonomy of each protein was labeled according to NCBI taxonomic assignment of the respective sequence. DIAMOND[265] (v0.9.14.115) was used to construct a protein database. Taxonomy is assigned by DIAMOND BLASTp[101] matches of proteins from an unknown phage sequence to the constructed database at the classifications of Order, Family and Sub-family. Assignment consists of reference protein taxonomy matching to each classification at the individual and all protein levels to hierarchically select the most likely taxonomic match rather than the most common (i.e., not recruitment of most

common match). Taxonomic assignments are available for 25 Families and 29 Sub-families for both bacterial and archaeal viruses. The database, script and associated files used to assign taxonomy are provided. To construct the protein network diagram vConTACT2[266] (v0.9.5, default parameters) was used to cluster mVCs with reference viruses from NCBI from the families *Ackermannviridae*, *Herelleviridae*, *Inoviridae*, *Microviridae*, *Myoviridae*, *Podoviridae* and *Siphoviridae* as well as several archaea-infecting families. The network was visualized using Cytoscape[267] (v3.7.2) and colored according to family affiliation.

Host prediction and CRISPR spacer analysis

A total of 25 representative metagenomes containing putative host sequences were downloaded from IMG (3300001676, 3300001678, 3300001679, 3300001680, 3300001681, 3300001683, 3300007516, 3300007722, 3300009154, 3300009182, 3300010157, 3300010296, 3300010297, 3300010299, 3300010300, 3300010318, 3300010354, 3300010370, 3300020258, 3300020264, 3300020266, 3300020314, 3300020325, 3300020365, 3300020454). Metagenome sequences were limited to a length of 10kb (149,986 total sequences). CRISPR Recognition Tool (CRT, v1.2, default settings)[268] was used to identify 7,178 CRISPR spacers from the 149,986 putative host sequences. Blastn (v2.2.31) was used to search the 191 mVC genome sequences for alignment to the spacers. A spacer hit was considered positive with 100% coverage to the spacer and 0-2 mismatches. To validate that the method worked properly, the 7,178 spacers were used to query the entire IMG/VR database (v2.1, October 2018).

World map distribution of mVCs

IMG/VR Taxon Object ID numbers respective of each mVCs were used to identify global coordinates of studies according to IMG documentation. Coordinates were mapped using Matplotlib (v3.0.0) Basemap[228] (v1.2.0). Human studies were excluded from coordinate maps.

Sequence alignments and conserved residues

Protein alignments were performed using MAFFT[214] (v7.388, default parameters). Visualization of alignments was done using Geneious Prime 2019.0.3. N- and C-terminal ends of protein alignments were manually removed, and aligned columns with 90% (SoxD and SoxYZ) or 98% (DsrA and DsrC/TusE) gaps were stripped (masked) for clarity. Amino acid residues were highlighted by pairwise identity of 90% (SoxC and SoxYZ) or 95% (DsrA, DsrC/TusE and SoxD). An identity graph, generated by Geneious, was fitted to the alignment to visualize pairwise identity of 100% (green), 99-30% (yellow) and 29-0% (red). Conservation of domains and amino acid residues was assessed according to annotations by The Protein Data Bank .

To calculate *dN/dS* ratios between mVC AMG pairs, dRep[219] (v2.6.2) was used to compare AMG sequences of *dsrA* (n = 39), *dsrC* (n = 141) and *soxYZ* (n = 44) separately (dRep compare --SkipMash --S_algorithm goANI). A custom auxiliary script (dnds_from_drep.py[269]) was used to calculate *dN/dS* ratios from the dRep output between various AMG pairs. Resulting *dN/dS* values were plotted using Seaborn[270] (v0.8.1) and Matplotlib. Phage AMG pairs and respective *dN/dS* values can be found in Supplementary Data 6.

mVC protein grouping

All protein sequences of 94 mVCs, excluding those with non-validated DsrC (i.e., potentially TusE-like) AMGs according to the conserved CxxxxxxxxxC motif, were grouped

using mmseqs2[271] (--min-seq-id 0.3 -c 0.6 -s 7.5 -e 0.001). For the AMG neighbor protein grouping, a total of 70 mVCs were used that encoded at least five proteins before and five proteins after the AMG. Groups containing at least two different representative mVCs were retained (887 groups total). A presence/absence heatmap was made using the R package "ComplexHeatmap"[272] and hierarchically grouped according to the ward.D method. Metadata for AMG, taxonomy and source environment were laid over the grouped columns. Two mVCs, Ga0066448_1000315 and JGI24724J26744_10000298, were not represented by any of the 887 retained clusters. mVC alignments were done using EasyFig[227] (v2.2.2).

To validate that our grouping method accurately depicts whole genome diversity of the mVCs, Caudovirales phages from the NCBI RefSeq database were used as a comparison. All Caudovirales phages were downloaded and dereplicated by 97% identity and 90% coverage using Mash and Nucmer. The dereplicated set consisted of 4413 RefSeq phages. A total of 94 RefSeq phages were randomly selected. Proteins were predicted using Prodigal and all proteins were grouped as before (mmseqs2 --min-seq-id 0.3 -c 0.6 -s 7.5 -e 0.001, minimum group size of 2 members). Random phage selection and protein grouping was performed 100 independent times (iterations). Over the 100 iterations the 94 randomly selected phages encoded 6821 to 10123 proteins (average 8357) and generated 727 to 1289 protein groups (average 989). The number of clusters per protein ranged from 0.088 to 0.149 (average 0.119). These statistics were similar to those seen from the mVCs, which generated 794 clusters from 6015 encoded proteins. The number of clusters per protein was 0.132. Therefore, despite encoding fewer proteins on average compared to RefSeq Caudovirales, the mVCs generated a comparable number of clusters per protein (Supplementary Data 7).

mVC genome structure and organization

mVCs representative of each AMG family were selected. Annotations were performed using VIBRANT and the best scoring annotation was used. Genomes were visualized using Geneious Prime and manually colored according to function.

AMG protein phylogenetic tree reconstruction

DSM protein sequences from reference prokaryotes were downloaded from NCBI nr database (accessed May 2019). The results were manually filtered for accurate annotations. The curated results were clustered by 70% sequence similarity using CD-HIT[216] (v4.7). These representative sequences from individual clusters were aligned with the corresponding mVC AMG protein sequences using MAFFT (default settings). Alignments were subjected to phylogenetic tree reconstruction using IQ-TREE[273] (v1.6.9) with the following settings: -m MFP -bb 100 -s -redo -mset WAG,LG,JTT,Dayhoff -mrate E,I,G,I+G -mfreq FU -wbtl ("LG+G4" was chosen as the best-fit tree reconstruction model). The environmental origin information of each mVC AMG was used to generate the stripe ring within the phylogenetic tree in the operation frame of iTOL[274] online server.

Metagenomic mapping and gene coverage ratio calculation

The metagenomic reads were first dereplicated by a custom Perl script and trimmed by Sickle[275] (v1.33, default settings). The QC-passed metagenomic reads were used to map against the collection of genes of investigated metagenomic assemblies by Bowtie2[225] (v2.3.4.1). The gene coverage for each gene was calculated by "jgi_summarize_bam_contig_depths" command within metaWRAP[96] (v1.0.2). The phage:total gene coverage ratio was calculated by adding up all the

phage and bacterial gene coverage values and using it to divide the summed phage gene coverage values.

We identified the phage-host gene pairs in the phylogenetic tree containing AMG and their bacterial counterpart gene encoding proteins. We assigned the phage-host gene pairs according to the following two criteria: 1) The phage and host gene encoding proteins are phylogenetically close in the tree; the branches containing them should be neighboring branches. 2) They should be from the same metagenomic dataset, which means that AMGs and bacterial host genes are from the same environment sample. The identified phage-host gene pairs were labelled accordingly in the phylogenetic tree.

For the gene coverage ratio calculation of phage genes and bacterial genes within a phage-host pair, we first calculated the phage:total gene coverage ratio and bacterial:total gene coverage ratio using the same method as described above; and then, in order to avoid the influence of numbers of phage or bacterial genes, we normalized the above two ratio values by the number of phage and bacterial genes, respectively. Finally, the normalized phage:host gene coverage ratio of this phage-host pair was calculated by comparing these two ratio values, accordingly.

Additionally, reads mapping performance was re-checked by comparing original mapping results (using Bowtie 2 "-very-sensitive" option) to the mapping results that only include reads with one mismatch (Supplementary Fig. 8). Checking results have justified the reliability of our original mapping performance and our gene coverage ratio calculation.

Metatranscriptomic mapping

The metatranscriptomic reads were first dereplicated by a custom Perl script and trimmed by Sickle (default settings), and then subjected to rRNA-filtering using SortMeRNA[276] (v2.0) with

the 8 default rRNA databases (including prokaryotic 16S rRNA, 23S rRNA; eukaryotic 18S rRNA, 28S rRNA; and Rfam 5S rRNA and 5.8S rRNA). QC-passed metagenomic reads were mapped against the collection of AMGs using Bowtie2 (--very-sensitive). The gene expression level in Reads Per Kilobase per Million mapped reads (RPKM) was calculated by normalizing the sequence depth (per million reads) and the length of the gene (in kilobases).

Data Availability

All IMG/VR sequences are available at https://img.jgi.doe.gov/cgi-bin/vr/main.cgi and https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=IMG_VR. Sequences from identified mVCs are available publicly and described in Supplementary Data 1 and 2. Any other relevant data are available from the authors upon request. All sequences and custom analysis scripts used in this study are also available at https://github.com/AnantharamanLab/Kieft_and_Zhou_et_al._2020.

Acknowledgements

We thank the University of Wisconsin—Office of the Vice Chancellor for Research and Graduate Education, University of Wisconsin—Department of Bacteriology, and University of Wisconsin—College of Agriculture and Life Sciences for their support. K.K. is supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison. This work was partly supported by the National Science Foundation grants OCE-2049478 to KA and OCE-0961947 to MBS. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231.

Author Contributions

K.K., Z.Z., S.R. and K.A. designed the study. K.K. and S.R. identified the genomes. K.K., Z.Z., and K.A. conducted the analyses. K.K., Z.Z., and K.A. drafted the manuscript. All authors (K.K., Z.Z., R.E.A., A.B., B.J.C., S.J.H., M.H., M.B.S., D.A.W., S.R., and K.A.) reviewed the results, revised and approved the manuscript.

Chapter 4: VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences

Kristopher Kieft[1], Zhichao Zhou[1], and Karthik Anantharaman[1]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA

Publication:
Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).

All supplementary figures, tables, and files are available at the following Figshare repository:
https://figshare.com/projects/Kristopher_Kieft_PhD_Dissertation/136427

Abstract

Background

Viruses are central to microbial community structure in all environments. The ability to generate large metagenomic assemblies of mixed microbial and viral sequences provides the opportunity to tease apart complex microbiome dynamics, but these analyses are currently limited by the tools available for analyses of viral genomes and assessing their metabolic impacts on microbiomes.

Design

Here we present VIBRANT, the first method to utilize a hybrid machine learning and protein similarity approach that is not reliant on sequence features for automated recovery and annotation of viruses, determination of genome quality and completeness, and characterization of viral community function from metagenomic assemblies. VIBRANT uses neural networks of protein signatures and a newly developed v-score metric that circumvents traditional boundaries to maximize identification of lytic viral genomes and integrated proviruses, including highly diverse viruses. VIBRANT highlights viral auxiliary metabolic genes and metabolic pathways, thereby serving as a user-friendly platform for evaluating viral community function. VIBRANT was trained and validated on reference virus datasets as well as microbiome and virome data.

Results

VIBRANT showed superior performance in recovering higher quality viruses and concurrently reduced the false identification of non-viral genome fragments in comparison to other virus identification programs, specifically VirSorter, VirFinder and MARVEL. When applied to 120,834 metagenomically derived viral sequences representing several human and natural environments, VIBRANT recovered an average of 94% of the viruses, whereas VirFinder, VirSorter and MARVEL achieved less powerful performance, averaging 48%, 87% and 71%,

respectively. Similarly, VIBRANT identified more total viral sequence and proteins when applied to real metagenomes. When compared to PHASTER, Prophage Hunter and VirSorter for the ability to extract integrated provirus regions from host scaffolds, VIBRANT performed comparably and even identified proviruses that the other programs did not. To demonstrate applications of VIBRANT, we studied viromes associated with Crohn's Disease to show that specific viral groups, namely Enterobacteriales-like viruses, as well as putative dysbiosis associated viral proteins are more abundant compared to healthy individuals, providing a possible viral link to maintenance of diseased states.

Conclusions

The ability to accurately recover viruses and explore viral impacts on microbial community metabolism will greatly advance our understanding of microbiomes, host-microbe interactions and ecosystem dynamics.

Background

Viruses that infect bacteria and archaea are globally abundant, and outnumber their hosts in most environments [277–279]. Viruses are obligate intracellular pathogenic genetic elements capable of reprogramming host cellular metabolic states during infection and can cause the lysis of 20-40% of microorganisms in diverse environments every day [37,57]. Due to their abundance and widespread activity, viruses are key facets in microbial communities as they contribute to cycling of essential nutrients such as carbon, nitrogen, phosphorus and sulfur [5,48,136,137,280]. In human systems, viruses have been implicated in contributing to dysbiosis that can lead to various diseases, such as inflammatory bowel diseases, or even have a symbiotic role with the immune system [21,147,281].

Viruses harbor vast potential for diverse genetic content, arrangement and encoded functions [2,40,41,45]. Recognizing their genetic diversity, there has been substantial interest in "mining" these viral sequences for novel anti-microbial drug candidates, enzymes for biotechnological applications, and for bioremediation [32,33,282–284]. Recently, it has been appreciated that viruses may directly link biogeochemical cycling of nutrients by specifically driving metabolic processes [73,74,82,84,85]. For example, during infection viruses can acquire 40-90% of their required nutrients from the surrounding environment by taking over and subsequently directing host metabolism [66,68,69]. To manipulate host metabolic frameworks, some viruses selectively "steal" metabolic genes from their host. These host derived genes, collectively termed auxiliary metabolic genes (AMGs), can be actively expressed during infection to provide viruses with fitness advantages [38,44,75,140]. Due to the need to study the role of viruses in microbiomes and integrate viruses into models of ecosystem function, it has become of great interest to determine which sequences within whole microbial communities are derived from viruses. These sequences can include free virions, active intracellular infections (which may be the case for as many as 30% of all bacteria at any given time [285]), particle or host-attached virions [286], and host-integrated or episomal viral genomes (i.e., proviruses).

Multiple tools exist for the identification of viruses from mixed metagenomic assemblies. For several years VirSorter [115], which succeeded tools such as VIROME [287] and Metavir [288], has been the most widely used for its ability to identify viral metagenomic fragments (scaffolds) from large metagenomic assemblies. VirSorter predominantly relies on database searches of predicted proteins, using both reference homology as well as probabilistic similarity, to compile metrics of enrichment of virus-like proteins and simultaneous depletion of other proteins. To do this it uses a virus-specific curated database as well as Pfam [289] for non-virus annotations, though it does not

fully differentiate viral from non-viral Pfam annotations. It also incorporates sequence signatures of viral genomes, such as encoding short genes or having low levels of strand switching between genes. VirSorter is also unique in its ability to use these annotation and sequence metrics to identify and extract integrated provirus regions from host scaffolds.

More recent tools have been developed as alternatives or supplements of VirSorter. VirFinder [113] was the first tool to implement machine learning and be completely independent of annotation databases for predicting viruses, which was a platform later implemented in PPR-Meta [290]. VirFinder was built with the consideration that viruses tend to display distinctive patterns of 8-nucleotide frequencies (otherwise known as 8-mers), which was proposed despite the knowledge that viruses can share remarkably similar nucleotide patterns with their host [291]. These 8-mer patterns are used to quickly classify sequences as short as 500 bp and generate model-derived scores, though it is up to the user to define the score cutoffs. VirFinder was shown to greatly improve the ability to recover viruses compared to VirSorter, but it also demonstrated substantial host and source environment biases in predicting diverse viruses, likely due to reference database-associated biases while training the machine learning model [113]. Moreover, under-recovery of viruses from certain environments was identified [292].

Additional recent tools have been developed that utilize slightly different methods for identifying viruses. MARVEL [293], for example, leverages annotation, sequence signatures and machine learning to identify viruses from metagenomic bins. MARVEL differs from VirSorter in that it only utilizes a single virus-specific database for annotation. However, MARVEL provides no consideration for integrated proviruses and is only suitable for identifying bacterial dsDNA viruses from the order *Caudovirales* which substantially limits its ability to discover novel viruses. Another recently developed tool, VirMiner [294], is unique in that it functions to use metagenomic

reads and associated assembly data to identify viruses and performs best for highly abundant viruses. VirMiner is a web-based server that utilizes a hybrid approach of employing both homology-based searches to a virus-specific database as well as machine learning. VirMiner was found to have improved ability to recover viruses compared to both VirSorter and VirFinder but was concurrently much less accurate.

Thus far, VirSorter remains the most efficient tool for identifying integrated proviruses within metagenomic assemblies. Other tools, predominantly PHASTER [295] and Prophage Hunter [296], are specialized in identifying integrated proviruses from whole genomes rather than scaffolds generated by metagenomic assemblies. Similar to VirSorter, these two provirus predictors rely on reference homology and viral sequence signatures with sliding windows to identify regions of a host genome that belong to a virus. Although they are useful for whole genomes, they lack the capability of identifying scaffolds belonging to lytic (i.e., non-integrated) viruses and perform slower for large datasets. In addition, both PHASTER and Prophage Hunter are exclusively available as web-based servers and offer no stand-alone command line tools.

Here we developed VIBRANT (Virus Identification By iteRative ANnoTation), a tool for automated recovery, annotation, and curation of both free and integrated viruses from metagenomic assemblies and genome sequences. VIBRANT is capable of identifying diverse dsDNA, ssDNA and RNA viruses infecting both bacteria and archaea, and to our knowledge has no evident environmental biases. VIBRANT uses neural networks of protein annotation signatures from non-reference-based similarity searches with Hidden Markov Models (HMMs) as well as a unique ‘v-score’ metric to maximize identification of diverse and novel viruses. After identifying viruses VIBRANT implements curation steps to validate predictions. VIBRANT additionally characterizes viral community function by highlighting AMGs and assesses the metabolic

pathways present in viral communities. All viral genomes, proteins, annotations and metabolic profiles are compiled into formats for user-friendly downstream analyses and visualization. When applied to reference viruses, non-reference virus datasets and various assembled metagenomes, VIBRANT outperformed VirFinder, VirSorter and MARVEL in the ability to maximize virus recovery and minimize false discovery. When compared to PHASTER, Prophage Hunter and VirSorter for the ability to extract integrated provirus regions from host scaffolds, VIBRANT performed comparably. VIBRANT was also used to identify differences in metabolic capabilities between viruses originating from various environments. When applied to three separate cohorts of individuals with Crohn's Disease, VIBRANT was able to identify both differentially abundant viral groups compared to healthy controls as well as virally encoded genes putatively influencing a diseased state. VIBRANT is freely available for download at https://github.com/AnantharamanLab/VIBRANT. VIBRANT is also available as a user-friendly, web-based application through the CyVerse Discovery Environment at https://de.cyverse.org/de [297].

Methods

Dataset for generation and comparison of metrics

To generate training and testing datasets, sequences representing bacteria, archaea, plasmids and viruses were downloaded from the National Center for Biotechnology Information (NCBI) RefSeq and Genbank databases (accessed July 2019) (Additional File 1: Table S1). For bacteria/archaea, 181 genomes were chosen by selecting from diverse phylogenetic groups. Likewise, a total of 1,452 bacterial plasmids were chosen. For viruses, NCBI taxids associated with viruses that infect bacteria or archaea were used to download reference virus genomes, which

were then limited to only sequences above 3kb. This included viruses with both DNA and RNA genomes, though RNA genomes must first be converted to complementary DNA. Sequences not associated with genomes, such as partial genomic regions, were identified according to sequence headers and removed. This resulted in 15,238 total viral partial and complete genomes. To be consistent between all sequences acquired from NCBI, proteins and genes were predicted using Prodigal (-p meta, v2.6.3) [98]. All sequences were split into non-overlapping, non-redundant fragments between 3kb and 15kb to simulate metagenome assembled scaffolds. These simulated scaffolds are hereafter called *fragments* and were used throughout training and testing VIBRANT. For RNA virus detection 33 viral (bacteriophage) genomes from NCBI RefSeq and 37 from Krishnamurthy *et. al*. were used [43], and for archaeal virus detection all genomes were acquired from NCBI RefSeq. The RNA and archaeal viral genomes were represented in both the training and testing datasets as genomic fragments and recall evaluation was performed on whole genomes. These were the only datasets in which training and evaluation datasets were semi-redundant. See Supplemental Methods (Additional File 16) for additional datasets and sequences used.

Integrated viruses are common in both bacteria and archaea. To address this for generating a dataset devoid of viruses, PHASTER (accessed July 2019) was used to predict putative integrated viruses in the 181 bacteria/archaea genomes. Using BLASTn [101], any fragments that had significant similarity (at least 95% identity, at least 3kb coverage and e-value < 1e-10) to the PHASTER predictions were removed as contaminant virus sequence. The new bacteria/archaea dataset was considered depleted of proviruses, but not entirely devoid of contamination. Next, the datasets for bacteria/archaea and plasmids were annotated with KEGG, Pfam and VOG HMMs (hmmsearch (v3.1), e-value < 1e-5) [102] to further remove contaminant virus sequence (see next section for details of HMMs). Plasmids were included because it was noted that the dataset appeared to

contain virus sequences, possibly due to misclassification of episomal proviruses as plasmids. Using manual inspection of the KEGG, Pfam and VOG annotations any sequence that clearly belonged to a virus was removed. Manual inspection was guided first by the number of KEGG, Pfam and VOG annotations, and then by the annotations themselves. For example, sequences with more VOG than KEGG or Pfam annotations were inspected and removed if multiple viral hallmark genes were found or if the majority of annotations represented viral-like genes. The final datasets consisted of 400,291 fragments for bacteria/archaea, 14,739 for plasmids, and 111,963 for viruses. Total number of fragments for all datasets used can be found in Additional File 2: Table S2.

Databases used by VIBRANT

VIBRANT uses HMM profiles from three different databases: Kyoto Encyclopedia of Genes and Genomes (KEGG) KoFam (March 2019 release) [212,298], Pfam (v32) [289] and Virus Orthologous Groups (VOG) (release 94, vogdb.org). For Pfam all HMM profiles were used. To increase speed, KEGG and VOG HMM databases were reduced in size to contain only profiles likely to annotate the viruses of interest. For KEGG this was done by only retaining profiles considered to be relevant to "prokaryotes" as determined by KEGG documentation. For VOG this was done by only retaining profiles that had at least one significant hit to any of the 15,238 NCBI-acquired viruses using BLASTp. The resulting databases consisted of 10,033 HMM profiles for KEGG, 17,929 for Pfam, and 19,182 for VOG (Additional File 3: Table S3).

V-score generation

Predicted proteins from reference viral genomes from NCBI and VOG database viral proteins were combined to generate v-scores, which resulted in a total of 633,194 proteins.

Redundancy was removed from the viral protein dataset using CD-HIT (v4.6) [216] with a identify cutoff of 95%, which resulted in a total of 240,728 viral proteins. This was the final dataset used to generate v-scores. All KEGG HMM profiles were used to annotate the viral proteins. A v-score for each KEGG HMM profile was determined by the number of significant (e-value < 1e-5) hits by hmmsearch, divided by 100, and a maximum value was set at 10 after division. The same v-score generation was done for Pfam and VOG databases. Any HMM profile with no significant hits to the virus dataset was given a v-score of zero. For KEGG and Pfam databases, any annotation that was given a v-score above zero and contained the keyword "phage" was given a minimum v-score of 1. To highlight viral hallmark genes, any annotation within all three databases with the keyword *portal, terminase, spike, capsid, sheath, tail, coat, virion, lysin, holin, base plate, lysozyme, head* or *structural* was given a minimum v-score of 1. Non-prokaryotic virus annotations (e.g., *reovirus core-spike protein*) were not considered. Each HMM is assigned a v-score and represents a metric of virus association (i.e., do not take into account virus specificity, or association with non-viruses) and are manually tuned to put greater weight on viral hallmark genes (Additional File 4: Table S4). Overall, annotations that are likely non-viral will have a low v-score whereas annotations that are commonly associated with viruses will have a high v-score. Raw HMM table outputs for v-score generation can be found in Additional Files 5, 6 and 7 for KEGG, Pfam and VOG, respectively (Additional File 5: Table S5, Additional File 6: Table S6 and Additional File 7: Table S7).

Training and testing VIBRANT

The bacteria/archaea genomic, plasmid and virus datasets described above were used to train and test the machine learning model. Scikit-Learn (v0.21.3) [299] libraries were used to assess

various machine learning strategies to identify the best performing algorithm. Among support vector machines, neural networks and random forests, we found that neural networks lead to the most accurate and comprehensive identification of viruses. Therefore, Scikit-Learn's supervised neural network multi-layer perceptron classifier (hereafter called neural network) was used. The portion of VIBRANT's workflow up until the neural network classifier (i.e., KEGG, Pfam and VOG annotation) was used to compile the 27 annotation metrics for each scaffold. To account for differences in scaffold sizes all metrics are normalized (i.e., divided by) to the total number of proteins encoded by the scaffold. The first metric, for total proteins, was normalized to log base 10 of itself. Each metric was weighted equally, though it is worth noting that the removal of several metrics did not significantly impact the accuracy of model's prediction. The normalized results were randomized, and non-redundant portions of these results were taken for training or testing the neural network. In total, 93,913 fragments were used for training and 9,000 different fragments were used for testing the neural network specifically (Additional File 8: Table S8 and Additional File 9: Table S9).

To test the performance of VIBRANT in its entirety, a new testing dataset was generated consisting of fragments from the neural network testing set as well as additional fragments non-redundant to the previous training dataset (hereafter called comprehensive test dataset). This new comprehensive test dataset was comprised of 256,713 genomic fragments from bacteria/archaea, 29,926 from viruses and 8,968 from plasmids. Each met the minimum protein number requirement of VIBRANT: at least four open reading frames.

Calculation of evaluation metrics and benchmarking of VIBRANT

For comparison of VIBRANT (v1.2.0) to VirFinder (v1.1), VirSorter (v1.0.3) and MARVEL (v0.2), the comprehensive test dataset was used. Two intervals for VirFinder and VirSorter were used for comparison. For VirSorter, the intervals selected were (1) category 1 and 2 predictions, and (2) category 1 and 2 predictions using the *virome decontamination mode*. Categories 1 and 2 are generally considered trustworthy, but category 3 predictions are more likely to contain false identifications. VirSorter was ran using the "Virome" database. For VirFinder, the intervals were (1) scores greater than or equal to 0.90 (approximately equivalent to a p-value of 0.013), and (2) scores greater than or equal to 0.75 (approximately equivalent to a p-value of 0.037). Since MARVEL was built for the identification of viral bins, each scaffold was evaluated separately as a single "bin". To ensure proper identification by MARVEL and VIBRANT, different versions of Scikit-Learn were used for each (v0.19.1 and v0.21.3, respectively).

Several metrics were used to compare performance of all four programs: recall, precision, accuracy, specificity, Mathews Correlation Coefficient (MCC) and F1 score. When calculating metrics, the larger bacteria/archaea and plasmid dataset was normalized to the size of the smaller viral dataset in order to make accurate calculations. All equations used can be found in Additional File 10: Table S10 and the results of each calculation can be found in Additional File 11: Table S11. Comparison metrics were visualized using R (v3.5.2) package "ggplot2".

It is worth noting that although VIBRANT was tested using sequences that were not used for training, biases may still be associated with reported metrics due to the reliance of KEGG, Pfam and VOG HMMs on NCBI databases. That is, NCBI databases in part were used to construct the HMMs and therefore are well suited at annotating NCBI-derived sequences. This same type of bias will be seen in the evaluation of VirSorter and MARVEL, both of which rely on NCBI-reliant databases. Although VirFinder does not use annotation databases, the machine learning algorithm

it employs was trained on NCBI-derived sequences. Similarly, biases with comparisons to VirFinder, VirSorter and MARVEL will arise when using NCBI databases. Sequences from NCBI were used for training each of the three programs and therefore will likely contain redundancy to VIBRANT's comprehensive test dataset. This redundancy will cause artificially enhanced performance. To address these biases, we further compared all four programs to non-NCBI datasets (see below).

AMG identification

KEGG annotations were used to classify potential AMGs (Additional File 12: Table S12). KEGG annotations falling under the "metabolic pathways" category as well as "sulfur relay system" were considered. Manual inspection was used to remove non-AMG annotations, such as *nrdAB* and *thyAX*. Other annotations not considered were associated with direct nucleotide to nucleotide conversions. All AMGs were associated with a KEGG metabolic pathway map.

Completeness estimation

Scaffold completeness is determined based on four metrics: circularization of scaffold sequence, VOG annotations, total VOG nucleotide replication proteins and total VOG viral hallmark proteins (Additional File 13: Table S13). In order to be considered a complete genome a sequence must be identified as likely circular. A kmer-based approach is used to do this. Specifically, the first 20 nucleotides are compared to 20-mer sliding windows within the last 900bp of the sequence. If a complete match is identified the sequence is considered a circular template. Scaffolds can also be considered a low, medium or high-quality draft. To benchmark completeness, 2466 NCBI RefSeq viruses identified as *Caudovirales*, limited to 10 kb in length,

were used to estimate completeness by stepwise removing 10% viral sequence at a time. VIBRANT was found to identify 2465 of the 2466 viruses. This set of viruses was additionally used to assess the error rate of cutting provirus regions. Viral genome diagrams to depict genome quality and completeness, provirus predictions and novel virus identification, were made using Geneious Prime 2019.0.3.

Analysis of Crohn's Disease metagenomes

Metagenomic reads from He *et al.* [300] were assembled by Pasolli *et al.* [160] and used for analysis. VIBRANT (-l 5000) was used to predict viruses from 49 metagenomes originating from individuals with Crohn's Disease and 53 from healthy individuals (102 total samples). A total of 14,121 viruses were identified. Viral sequences were dereplicated using Mash [263] and Nucmer [301] to 95% nucleotide identity and 70% sequence coverage. The longest sequence was kept as the representative for a total of 8,822 dereplicated viruses. A total of 96 read sets were used (59 Crohn's Disease and 37 healthy), trimmed using Sickle and aligned to the dereplicated viruses using Bowtie2 (-N 1, v2.3.4.1) [225] and the resulting coverages were normalized to total reads. The normalized relative coverage of each virus for all 96 samples were compared using DESeq2 [302] (Additional File 14: Table S14). Viruses that displayed significantly different abundance between Crohn's Disease and control samples were determined by a p-value cutoff of 0.05. iRep (default parameters) [156] was used to estimate replication activity of two highly abundant Crohn's-associated viruses. EasyFig (v2.2.2) [227] was used to generate genome alignments of Escherichia phage Lambda (NCBI accession number NC_001416.1) and three Crohn's-associated viruses. vConTACT2 (v0.9.8) was run using default parameters on the CyVerse Discovery Environment platform. Putative hosts of Crohn's-associated and healthy-associated was estimated using

proximity of vConTACT2 protein clustering and BLASTp identity (NCBI non-redundant protein database, assessed October 2019). Two additional read sets from Gevers *et al.* [303] and Ijaz *et al.* [304] were likewise assembled by Pasolli *et al.*. VIBRANT (-l 5000 -o 10) was used to predict viruses from 43 metagenomes originating from individuals with Crohn's Disease and 21 from healthy individuals (64 total samples). In contrast to the discovery dataset viral genomes were not dereplicated and differential abundance was not determined. Instead viruses from each group were directly clustered using vConTACT2. Abundances of dysbiosis associated genes in the validation set were normalized to total viruses. Validation of dysbiosis associated genes' presence on viral genomes, rather than microbial contamination, was done by identifying viral hallmark genes on the viral scaffold (Additional File 15: Table S15). Protein networks were visualized using Cytoscape (v3.7.2) [267].

Results

VIBRANT was built to extract and analyze bacterial and archaeal viruses from assembled metagenomic and genome sequences, as well as provide a platform for characterizing metabolic proteins and functions in a comprehensive manner. The concept behind VIBRANT's mechanism of virus identification stems from the understanding that arduous manual inspection of annotated genomic sequences produces the most dependable results. As such, the primary metrics used to inform validated curation standards and to train VIBRANT's machine learning based neural network to identify viruses reflects human-guided intuition, though in a high-throughput automated fashion.

Determination of v-score

We developed a unique 'v-score' metric as an approach for providing quantitative information to VIBRANT's algorithm in order to assess the qualitative nature of annotation information. A v-score is a value assigned to each possible protein annotation that scores its association to known viral genomes (see Methods). V-score differs from the previously used "virus quotient" metric [305,306] in that it does not take into account the annotation's relatedness to bacteria or archaea. Not including significant similarity to non-viral genomes in the calculation of v-scores has important implications for this metric's utility. Foremost is that annotations shared between viruses and their hosts, such as ribonucleotide reductases, will be assigned a v-score reflecting its association to viruses, not necessarily virus-specificity. Many genes are commonly associated with viruses and host organisms, but when encoded on viral genomes can be central to virus replication efficiency (e.g., ribonucleotide reductases [307]). Therefore, a metric representing virus-association rather than virus-specificity would be more appropriate in identifying if an unknown scaffold is viral or not. Secondly, this approach takes into account widespread horizontal gene transfer of host genes by viruses as well as the presence of AMGs.

VIBRANT workflow

VIBRANT utilizes several annotation metrics in order to guide removal of non-viral scaffolds before curation of reliable viral scaffolds. The annotation metrics used are derived from HMM-based probabilistic searches of protein families from the KEGG, Pfam and VOG databases. VIBRANT is not reliant on reference-based similarity and therefore accounts for the large diversity of viruses on Earth and their respective proteins. Consequently, widespread horizontal gene transfer, rapid mutation and the vast amount of novel sequences do not hinder VIBRANT's ability to identify known and novel viruses. VIBRANT does not rely on non-annotation features, such as

rates of open reading frame strand switching, because these features were not as well conserved in genomic scaffolds in contrast to whole genomes.

VIBRANT's workflow consists of four main steps (Figure 1A). Briefly, proteins (predicted or user input) are used by VIBRANT to first eliminate non-viral sequences by assessing non-viral annotation signatures derived from KEGG and Pfam HMM annotations. At this step potential host scaffolds are fragmented using sliding windows of KEGG annotation v-scores in order to extract integrated provirus sequences. Following the elimination of most non-viral scaffolds and rough excision of provirus regions, proteins are annotated by VOG HMMs. Before analysis by the neural network machine learning model, any extracted putative provirus is trimmed to exclude any remaining non-viral sequences. Annotations from KEGG, Pfam and VOG are used to compile 27 metrics that are utilized by the neural network to predict viral sequences (Additional File 16: Supplemental Methods). These 27 metrics were found to be adequate for the separation of viral and non-viral scaffolds (Figure 1B).

After prediction by the neural network a set of curation steps are used to filter the results. Curation is an automated mechanism of verifying and/or altering the neural network predictions in order to improve accuracy and recovery of viruses. This concept, as previously stated, originates from experiences with manual inspection of viral genomes that cannot be captured even within machine learning algorithms. For example, these curation steps can: (1) more accurately separate plasmid sequences by discerning viral-like and plasmid-like integrase annotations, (2) remove scaffolds that encode a high density of bacterial-like (i.e., v-score of zero) proteins, or (3) increase true positive identifications by retaining otherwise missed scaffolds that are unique (e.g., encode few but highly virus-related proteins).

Once viruses are identified VIBRANT automates the analysis of viral community function by highlighting AMGs and assigning them to KEGG metabolic pathways. The genome quality (i.e., proxy of completeness) of identified viruses is estimated using a subset of the annotation metrics and viral sequences are used to identify circular templates (i.e., likely complete circular viruses). These quality analyses were determined to best reflect established completeness metrics for both bacteria and viruses [308,309]. Finally, VIBRANT compiles all results into a user-friendly format for visualization and downstream analysis. For a detailed description of VIBRANT's workflow see Methods.

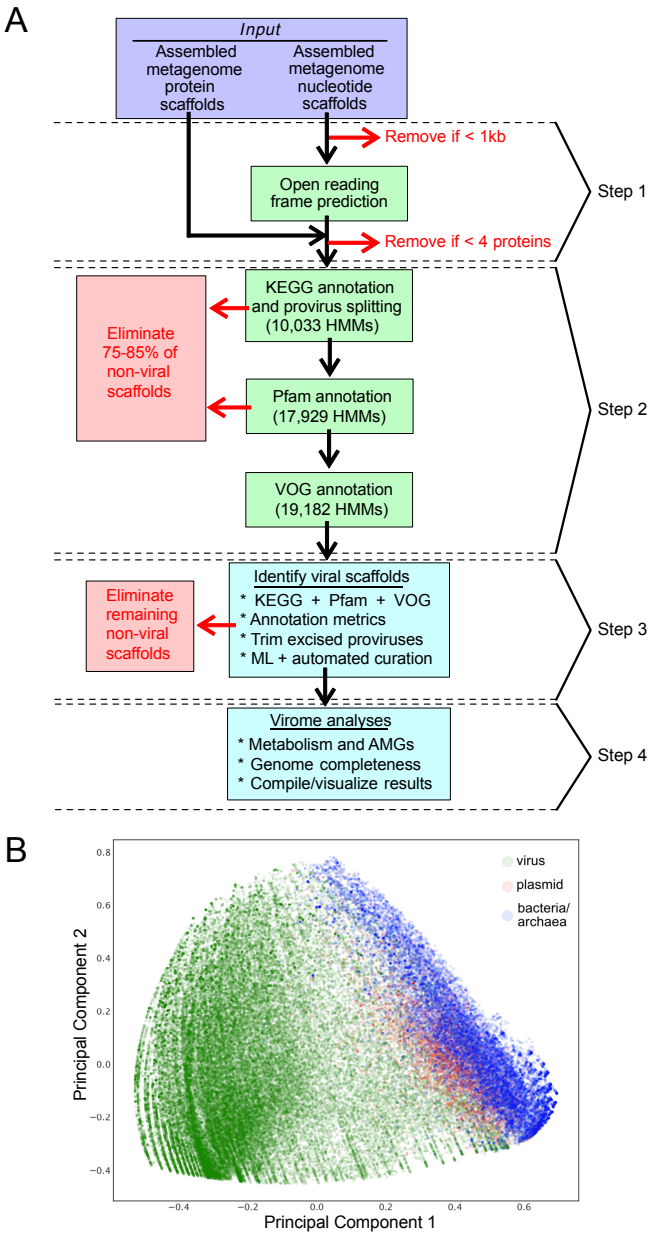


Figure 1. Representation of VIBRANT's method for virus identification and virome functional characterization. (A) Workflow of virome analysis. Annotations from KEGG, Pfam and VOG databases are used to construct signatures of viral and non-viral annotation signatures that are read into a neural network machine learning model. (B) Visual representation (PCA plot) of the metrics used by the neural network to identify viruses, depicting viral, plasmid and bacterial/archaeal genomic sequences.

Comparison of VIBRANT to other programs

VirSorter, VirFinder and MARVEL, three commonly used programs for identifying bacterial and archaeal viruses from metagenomes, were selected to compare against VIBRANT for the ability to accurately identify viruses. We evaluated all four programs' performance on the same viral, bacterial and archaeal genomic, and plasmid datasets. Given that both VirSorter and VirFinder produce various confidence ranges of virus identification, we selected certain parameters for each program for comparison. For VirSorter, the parameters selected were (1) category 1 and 2 predictions, and (2) category 1 and 2 predictions using the *virome decontamination mode*. For VirFinder, the intervals were (1) scores greater than or equal to 0.90 (approximately equivalent to a p-value of 0.013), and (2) scores greater than or equal to 0.75 (approximately equivalent to a p-value of 0.037). Hereafter, we provide two statistics for each VirSorter and VirFinder run that reflect results according to the two set confidence intervals, respectively. Both VIBRANT and MARVEL have set output predictions and therefore will be reported with a single statistic.

VIBRANT yields a single output of confident predictions and therefore does not provide multiple output options. Since VIBRANT is only partially reliant on its neural network machine learning model for making predictions, all comparisons are focused on VIBRANT's full workflow performance. VIBRANT does not consider scaffolds shorter than 1000 bp or those that encode less than four predicted open reading frames in order to maintain a low false positive rate (FPR) and have sufficient annotation information for identifying viruses. Therefore, in comparison of performance metrics only scaffolds meeting VIBRANT's minimum requirements were analyzed. Inclusion of fragments encoding less than four open reading frames in analyses, which are frequently generated by metagenomic assemblies, are discussed below. We used the following

statistics to compare performance: recall, precision, accuracy, specificity, MCC and F1 score (Figure 2).

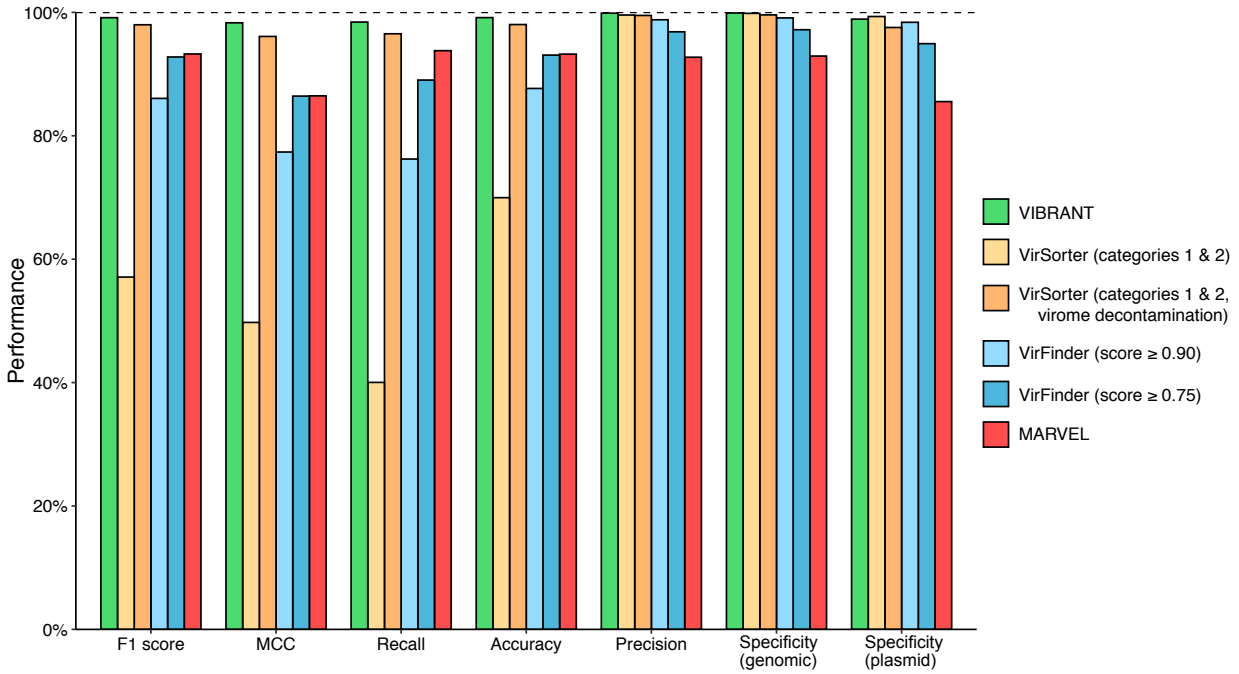


Figure 2. Performance comparison of VIBRANT, VirFinder, VirSorter and MARVEL on artificial scaffolds 3kb-15kb. Performance was evaluated using datasets of reference viruses, bacterial plasmids, and bacterial/archaeal genomes. For VirFinder and VirSorter two different confidence cutoffs were used (VirFinder: score of at least 0.90, and score of at least 0.75. VirSorter: categories 1 and 2 predictions, and categories 1 and 2 predictions using virome decontamination mode). All four programs were compared using the following statistical metrics: F1 score, MCC, recall, precision, accuracy and specificity. To ensure equal comparison all scaffolds tested encoded at least four open reading frames.

First, we evaluated the true positive rate (TPR, or recall) of viral genomic fragments as well as whole viral genomes. Viral genomes were acquired from the NCBI RefSeq and GenBank databases and split into various non-redundant fragments between 3 and 15 kb to simulate genomic scaffolds (see Methods). VIBRANT correctly identified 98.43% of the 29,926 viral fragments, which was greater than VirSorter (40.03% and 96.53%), VirFinder (76.23% and 89.03%) and MARVEL (93.79%) at all scoring intervals. For VirSorter it was essential to set *virome decontamination mode* for datasets consisting of mainly viruses, without which the TPR was substantially inhibited.

Similar to TPR, we calculated FPR (or specificity) using two different datasets: genomic fragments of bacteria and archaea (hereafter called genomic), and bacterial plasmids (plasmid). Plasmids were evaluated separately because they often encode for genes similar to those on viral genomes, such as those for genome replication and mobilization. Genomic and plasmid sequences were acquired from NCBI RefSeq and GenBank databases and split into various non-redundant fragments between 3 and 15 kb and putative proviruses were depleted from the datasets (see Methods). VIBRANT had high specificity against both genomic (99.90%) and plasmid fragments (98.90%). VirSorter had similar specificity against both genomic (99.84% and 99.59%) and plasmid (99.33% and 97.55%) datasets, but only VirFinder set to a score cutoff of 0.90 was fully comparable (genomic: 99.10%, plasmid: 98.39%). VirFinder at a score cutoff of 0.75 (genomic: 97.19%, plasmid: 94.93%) along with MARVEL (genomic: 92.92%, plasmid: 85.54%) were slightly less specific. Although VirFinder (set to a score cutoff of 0.90) and VIBRANT had a similar overall specificity, VirFinder identified 9.3 times more genomic scaffolds as viruses (false discoveries) compared to VIBRANT (2,311 and 249, respectively). MARVEL was even more pronounced, identifying 72.9 times more genomic scaffolds as viruses (18,164 total) compared to VIBRANT.

We used the results from TPR of viral fragments and FPR of non-viral genomic or plasmid fragments to calculate precision (i.e., proportion of true virus identifications out of all virus identifications) and accuracy (i.e., proportion of correct predictions out of all predictions). VIBRANT outperformed each other program at both precision (VIBRANT: 99.87%, VirFinder: 98.80% and 96.85%, VirSorter: 99.57% and 99.50%, and MARVEL: 92.73%) and accuracy (VIBRANT: 99.15%, VirFinder: 87.67% and 93.08%, VirSorter: 69.97% and 98.03%, and MARVEL: 93.23%). F1 and MCC are additional metrics (maximum values of 1) accounting for

both TPR and FPR, and therefore acts as a comprehensive evaluation of overall performance. Our calculation of F1 indicates that VIBRANT (0.991) is able to better identify viruses while subsequently reducing false identifications compared to VirFinder (0.861 and 0.928), VirSorter (0.571 and 0.980) or MARVEL (0.933). MCC likewise indicated that VIBRANT (0.983) was better suited at maximizing the ratio of viruses to non-viruses compared to VirSorter (0.498 and 0.961), VirFinder (0.774 and 0.864) and MARVEL (0.865).

Although VIBRANT exhibits improved performance with scaffolds at least 3kb in length, it is worth noting that performance drops considerably at the set minimum length of 1kb. To display this, the TPR and FPR of both 1k and 3kb scaffolds were assessed (Additional File 16: Figure S1A). For this analysis, VirSorter was evaluated using virome decontamination mode and VirFinder was set to a score cutoff of 0.90. MARVEL's minimum length requirement is 2kb and therefore was not compared with 1kb scaffolds. For 1kb viral scaffolds, VIBRANT (1.95%) and VirSorter (1.12%) recovered far fewer scaffolds compared to VirFinder (22.56%). However, at a length of 3kb VIBRANT (43.54%) recovered more viral fragments than VirSorter (25.43%), VirFinder (34.42%) and MARVEL (37.82%). Even at the low resolution of short scaffolds VIBRANT's FPR is not impacted. For 1kb genomic and 1kb plasmid scaffolds VIBRANT (<0.00% and 0.07%) and VirSorter (<0.00% and 0.10%) had fewer false positive discoveries than VirFinder (2.61% and 3.70%). Similarly, for 3kb genomic and 3kb plasmid scaffolds VIBRANT (0.10% and 2.69%) and VirSorter (0.11% and 2.41%) falsely identified fewer sequences than VirFinder (2.26% and 5.54%) or MARVEL (6.08% and 16.30%). Overall, this suggests that VirFinder is uniquely able to accurately recover short (e.g., 1kb) viral scaffolds while maintaining a relatively low FPR, but this ability is not maintained with longer scaffolds. Moreover, our current abilities to sequence and assemble scaffolds of lengths over 3kb will likely lead to a greater focus

on longer viral sequences that are more amenable to downstream analysis, such as taxonomic classification and functional analyses.

Next, we assessed the ability of VIBRANT to filter out eukaryotic contamination rather than falsely identify these sequences as viral since eukaryotes were not represented in the training or testing datasets. However, these contaminants should be sparse because the majority of eukaryotic KEGG and VOG HMMs were removed from the annotation databases (see Methods). Likewise, eukaryotic-like annotations should receive a low v-score. A total of 8,672 eukaryotic sequences ranging from 1kb to 15kb were assessed. VIBRANT (0.62%), VirSorter (0.05% and 0.05%) and MARVEL (0.44%) performed well with recovering few sequences, whereas VirFinder (4.92% and 15.44%) recovered contamination at a greater rate (Additional File 16: Figure S1B).

Finally, viruses with RNA genomes as well as those that infect archaea are rare in current culture systems and sequence databases compared to bacterial dsDNA viruses. However, the true abundance of RNA and archaeal viruses has yet to be explored mainly due to biases towards dsDNA in genome extracting and sequencing methods [310] and the low abundance of archaea in most environments. VIBRANT was built to identify all prokaryotic viruses in order to expand our knowledge of understudied groups. A total of 70 RNA viral genomes and 93 archaeal viral genomes were used to evaluate recall. VIBRANT was able to recover 47% of RNA viruses, or 84% of the those that encoded at least four predicted open reading frames. In comparison, VirSorter (7% and 70%), VirFinder (33% and 57%) and MARVEL (68%) ranged from lower to higher recovery (Additional File 16: Figure S1C). The high recovery of RNA viruses by MARVEL is intriguing since the software was trained exclusively on dsDNA Caudovirales, but may be explained by the greater rate of false positive discovery. For archaeal viruses, VIBRANT (96.77%) identified significantly more viruses than VirSorter (70.97% and 93.55%), VirFinder (46.24% and

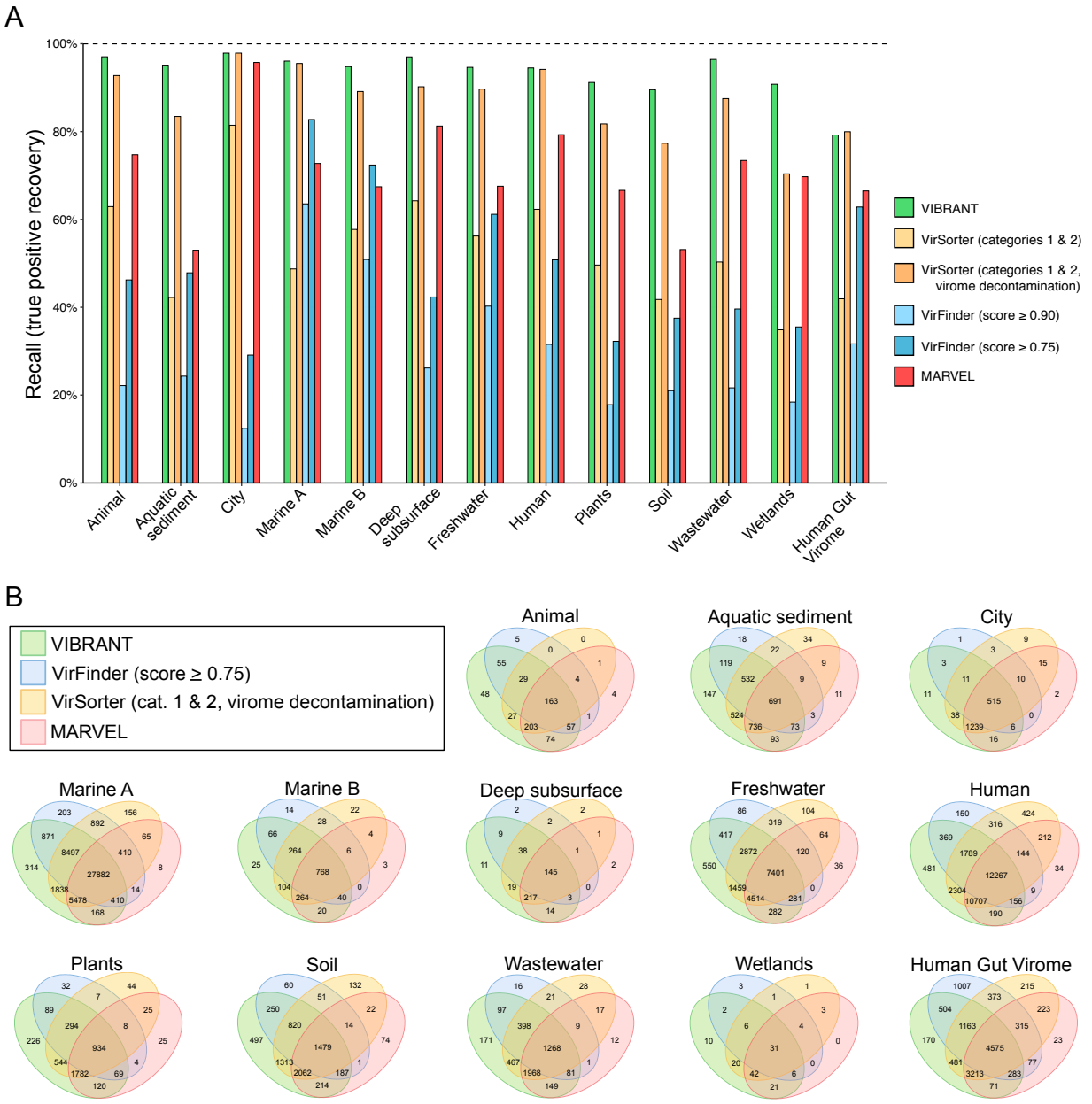


Figure 3. Effect of source environment on predictive abilities of VIBRANT, VirFinder, VirSorter and MARVEL. Viral scaffolds from IMG/VR and HGV database were used to test if VIBRANT displays biases associated with specific environments. (A) The recall (or recovery) of viral scaffolds from 12 environment groups was compared between VIBRANT and two confidence cutoffs for both VirFinder and VirSorter. Marine environments were classified into two groups: marine A (coastal, gulf, inlet, intertidal, neritic, oceanic, pelagic and strait) and marine B (hydrothermal vent, volcanic and oil). (B) Comparison of the overlap in the scaffolds identified as viruses by all three programs. Cutoffs for VirFinder (scores greater than or equal to 0.75) and VirSorter (categories 1 and 2 using virome decontamination mode) were set to display each program's ability to recover diverse viruses.

74.19%), and MARVEL (80.65%) (Additional File 16: Figure S1D). Taken together, VIBRANT has the potential to identify RNA and archaeal viruses, though the significance of this difference is hard to distinguish due to the current dearth of reference genomes with which to validate.

Identification of viruses in diverse environments

We next tested VIBRANT's ability to successfully identify viruses from a diversity of environments. Using 120,834 viruses from the IMG/VR database, in which the source environment of viruses is categorized, we identified that VIBRANT is more robust in identifying viruses from all tested environments compared to VirFinder, VirSorter and MARVEL (Figure 3A). The 12 environments were: animal-associated, aquatic sediment, city, marine A (coastal, gulf, inlet, intertidal, neritic, oceanic, pelagic and strait), marine B (hydrothermal vent, volcanic and oil), deep subsurface, freshwater, human-associated, plant-associated, soil, wastewater and wetlands. VIBRANT averaged 94.59% recall, substantially greater than VirFinder (29.19% and 48.13%), VirSorter (54.37% and 87.49%) and MARVEL (71.23%). Between the 12 environments VIBRANT recovered between 89.55% and 97.87% (total range of 8.33%) of the viruses. Conversely, VirFinder (score cutoff of 0.75) had a range of 53.65%, VirSorter (categories 1 and 2, virome decontamination) had a range of 27.48% and MARVEL had a range of 42.75%. These results suggest that in comparison to other software, VIBRANT has no evident environmental biases and is fully capable of identifying viruses from a broad range of source environments. We also used a dataset of 13,203 viruses from the Human Gut Virome database for additional comparison. The vast majority of viruses (~96%) in this dataset were assumed to infect bacteria. Although recall was diminished compared to IMG/VR datasets, VIBRANT (79.22%) nevertheless

outperformed or matched VirFinder (31.67% and 62.83%), VirSorter (41.93% and 79.97%) and MARVEL (66.49%) on this dataset.

Relatively few viruses from the IMG/VR dataset that were not identified by VIBRANT were identified by either VirFinder, VirSorter or MARVEL at even the most inclusive score cutoffs (Figure 3B). Furthermore, for most environments VIBRANT displayed the largest proportion of unique identifications, suggesting that VIBRANT has the propensity for discovery of viruses. The differences in the overlap of identified viruses was not too distinctive in environments for which many reference viruses are available, such as marine, though for more understudied environments, such as plants or wastewater, VIBRANT displayed near-complete overlap with VirFinder, VirSorter and MARVEL predictions. This suggests that database bias may not affect VIBRANT's performance to a significant degree. Although VirFinder does not rely on an annotation database, it still has been trained on a dataset of reference viral genomes which can contribute to database dependency and recall bias.

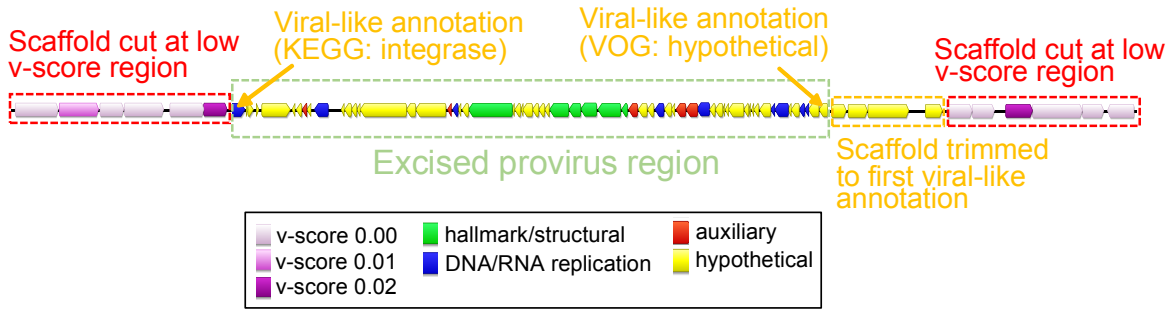
Identification of viruses in mixed metagenomes

Metagenomes assembled using short read technology contain many scaffolds that do not meet VIBRANT's minimum length requirements and therefore are not considered during analysis. Despite this, VIBRANT's predictions contain more annotation information and greater total viral sequence length than tools built to identify short sequences, such as scaffolds with less than four open reading frames. VIBRANT, VirFinder (score cutoff of 0.90) and VirSorter (categories 1 and 2) were used to identify viruses from human gut, freshwater lake and thermophilic compost metagenome sequences (Table 1). In addition, alternate program settings—VIBRANT *virome* mode, VirFinder at a score cutoff of 0.75 and VirSorter virome decontamination mode—were used

to identify viruses from an estuary virome dataset. MARVEL was not considered in this analysis due to the inability to achieve comparable precision. Each metagenomic assembly was limited to sequences of at least 1000bp but no minimum open reading frame limit was set. For these metagenomes, 31% to 40% of the scaffolds were of sufficient length (at least four open reading frames) to be analyzed by VIBRANT; for the estuary virome 62% were of sufficient length. In comparison, 100% of scaffolds from each dataset were long enough to be analyzed by VirFinder. The ability of VirFinder to make a prediction with each scaffold is considered the major strength of the tool.

For all six assemblies VirFinder averaged approximately 1.16 times more virus identifications than VIBRANT, though for both thermophilic compost and the estuary virome VIBRANT identified a greater number. Despite VirFinder averaging more total virus identifications, VIBRANT averaged 2.33 times more total viral sequence length and 2.44 times more total viral proteins. This is the result of VIBRANT having the capability to identify more viruses of higher quality and longer sequence length. For example, among all six datasets VIBRANT identified 1,320 total viruses at least 10 kb in length in comparison to VirFinder's 479. VIBRANT was also able to outperform VirSorter in all metrics, averaging 2.45 times more virus identifications, 1.76 times more total viral sequence length, and 1.86 times more encoded viral proteins. VIBRANT's method of predicting viruses provides a unique opportunity in comparison to similar tools in that it yields sequences of higher quality which are more amenable for analyzing protein function from virome data. It is an important distinction that the total number of viruses identified may not be correlated with the total viruses identified or the total number of encoded proteins. Even if VIBRANT identified fewer total viruses compared to other tools in certain circumstances, more data of higher quality was generated as viral sequences of longer length were

A



B

Genome	Region	VIBRANT	PHASTER	Prophage Hunter	VirSorter (cat. 1 & 2)
Lactococcus lactis	Putative provirus 1	35,516 - 55,047	28,459 - 56,369	34,907 - 49,863	10,280 - 95,312
	Putative provirus 2	442,048 - 483,244	447,143 - 484,064	447,143 - 484,094	450,421 - 517,314
	Putative provirus 3	492,452 - 531,873	502,338 - 520,485	-	-
	Putative provirus 4	1,033,815 - 1,070,779	1,033,815 - 1,079,173	1,018,719 - 1,075,596	1,029,339 - 1,080,357
	Putative provirus 5	1,416,048 - 1,460,426	1,414,112 - 1,457,044	1,415,289 - 1,457,116	1,414,112 - 1,460,426
	Putative provirus 6	-	1,997,699 - 2,028,023	2,011,247 - 2,025,756	2,011,426 - 2,025,635
Desulfovibrio vulgaris	Putative provirus 1	226,416 - 270,425	239,489 - 270,425	257,056 - 273,870	247,822 - 270,425
	Putative provirus 2	1,204,238 - 1,245,361	1,202,025 - 1,218,337	-	1,204,238 - 1,247,136
	Putative provirus 3	-	1,226,597 - 1,237,819	-	-
	Putative provirus 4	1,560,130 - 1,593,118	1,573,624 - 1,584,078	1,562,595 - 1,579,793	1,561,263 - 1,599,625
	Putative provirus 5	-	1,783,411 - 1,820,727	-	1,766,095 - 1,823,781
	Putative provirus 6	2,246,009 - 2,298,987	-	2,247,565 - 2,288,385	
	Putative provirus 7	2,715,796 - 2,730,860	-	-	2,709,005 - 2,749,853
	Putative provirus 8	2,799,591 - 2,834,336	2,803,529 - 2,834,336	-	2,799,591 - 2,834,336
	Putative provirus 9	2,936,283 - 2,971,476	2,936,186 - 2,977,658	-	2,936,283 - 2,979,715
Staphylococcus aureus	Putative provirus 1	1,245,539 - 1,251,661	1,242,808 - 1,251,661	1,240,995 - 1,253,838	-
	Putative provirus 2	1,452,300 - 1,508,497	1,451,176 - 1,523,261	1,460,593 - 1,511,405	1,462,371 - 1,514,191
	Putative provirus 3	-	-	1,812,933 - 1,834,606	-
	Putative provirus 4	1,923,883 - 1,968,119	1,907,832 - 1,966,929	1,910,863 - 1,977,294	1,922,968 - 1,966,929
	Putative provirus 5	2,034,690 - 2,073,429	2,029,598 - 2,077,786	2,031,642 - 2,080,909	2,031,741 - 2,074,574
Bacteroides vulgatus	Putative provirus 1	3,580,013 - 3,632,769	-	-	3,580,013 - 3,637,236

C

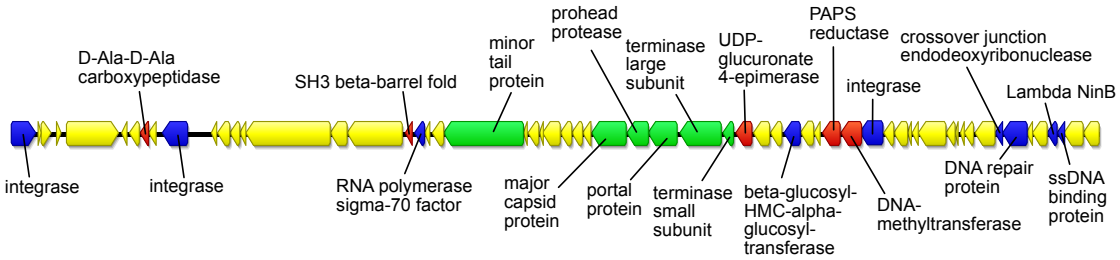


Figure 4. Prediction of integrated proviruses by VIBRANT, and comparison to PHASTER, Prophage Hunter and VirSorter. (A) Schematic representing the method used by VIBRANT to identify and extract provirus regions from host scaffolds using annotations. Briefly, v-scores are used to cut scaffolds at host-specific sites and fragments are trimmed to the nearest viral annotation. (B) Comparison of proviral predictions within four complete bacterial genomes between VIBRANT, PHASTER, Prophage Hunter and VirSorter. For PHASTER, putative proviruses are colored according to "incomplete" (red), "questionable" (blue) and "intact" (green) predictions. Prophage Hunter is colored according to "active" (green) and "ambiguous" (blue) predictions. All VirSorter predictions for categories 1 and 2 are shown in green. (C) Manual validation of the *Bacteroides vulgatus* provirus prediction made by VIBRANT. The presence of viral hallmark protein, integrase and genome replication proteins strongly suggests this is an accurate prediction.

identified as compared to many short fragments. This provides an important distinction that the metric of total viral predictions is not necessarily an accurate representation for the quality or quantity of the data generated.

Integrated provirus prediction

In many environments, integrated proviruses can account for a substantial portion of the active viral community [311]. Despite this, few tools exist that are capable of identifying both lytic viruses from metagenomic scaffolds as well as proviruses that are integrated into host genomes. To account for this important group of viruses, VIBRANT identifies provirus regions within metagenomic scaffolds or whole genomes. VIBRANT is unique from most provirus prediction tools in that it does not rely on sequence motifs, such as integration sites, and therefore is especially useful for partial metagenomic scaffolds in which neither the provirus nor host region is complete. In addition, this functionality of VIBRANT provides the ability to trim non-viral (i.e., host genome) ends from viral scaffolds. This results in a more correct interpretation of genes that are encoded by the virus and not those that are misidentified as being within the viral genome region. Briefly, VIBRANT identifies proviruses by first identifying and isolating scaffolds and genomes at regions spanning several annotations with low v-scores. These regions were found to be almost exclusive to host genomes. After cutting the original sequence at these regions, a refinement step trims the putative provirus fragment to the first instance of a virus-like annotation to remove leftover host sequence (Figure 4A). The final scaffold fragment is then analyzed by the neural network similar to non-excised scaffolds.

To assess VIBRANT's ability to accurately extract provirus regions we compared its performance to PHASTER and Prophage Hunter, two programs explicitly built for this task, as

well as VirSorter. We compared the performance of these programs with VIBRANT on four complete bacterial genomes. VIBRANT and PHASTER predicted an equal number of proviruses, 17, while Prophage Hunter and VirSorter identified slightly less with 13 and 16 identifications, respectively (Figure 4B). Only one putative provirus prediction (*Lactococcus lactis* putative provirus 6) was shared between all programs except VIBRANT. However, VIBRANT was able to identify two putative provirus regions (*Desulfovibrio vulgaris* putative provirus 7 and *Bacteroides vulgatus* putative provirus 1) that neither PHASTER nor Prophage Hunter identified, though VirSorter identified these likely due to the similar approach of extracting provirus regions. Manual inspection of the putative *Bacteroides vulgatus* provirus identified a number of virus hallmark and virus-like proteins suggesting that it is an accurate prediction (Figure 4C). Our results suggest VIBRANT has the ability to accurately identify proviruses and, in some cases, can outperform other tools in this task.

Both VIBRANT and VirSorter identify integrated proviruses from metagenomic assemblies by cutting host scaffolds at either end of a provirus region. By employing this method these programs generate a more comprehensive understanding of a virome, but errors in identified cut sites may occur due to the diversity of genomic arrangements in both virus and host. This will result in fragmented viral genomes that should have remained intact. We assessed the error rate of VIBRANT and VirSorter (using virome decontamination mode) for cutting viral genomes. A total of 2,466 *Caudovirales* complete genomes were acquired from the NCBI RefSeq database, including 74 megaphages with genomes greater than 200kb. In total, VIBRANT fragmented 5 genomes whereas VirSorter fragmented 159 (categories 1 and 2) or 160 (categories 1, 2 and 3). Although relatively comparable, VirSorter incorrectly cut 6.2% more complete viral genomes compared to VIBRANT (6.4% versus 0.2%, respectively).

Evaluating quality and completeness of predicted viral sequences

Determination of quality, in relation to completeness, of a predicted viral sequence has been notoriously difficult due to the absence of universally conserved viral genes. To date the most reliable metric of completeness for metagenomically assembled viruses is to identify circular sequences (i.e., complete circular genomes). Therefore, the remaining alternatives rely on estimation based on encoded proteins that function in central viral processes: replication of genomes and assembly of new viral particles.

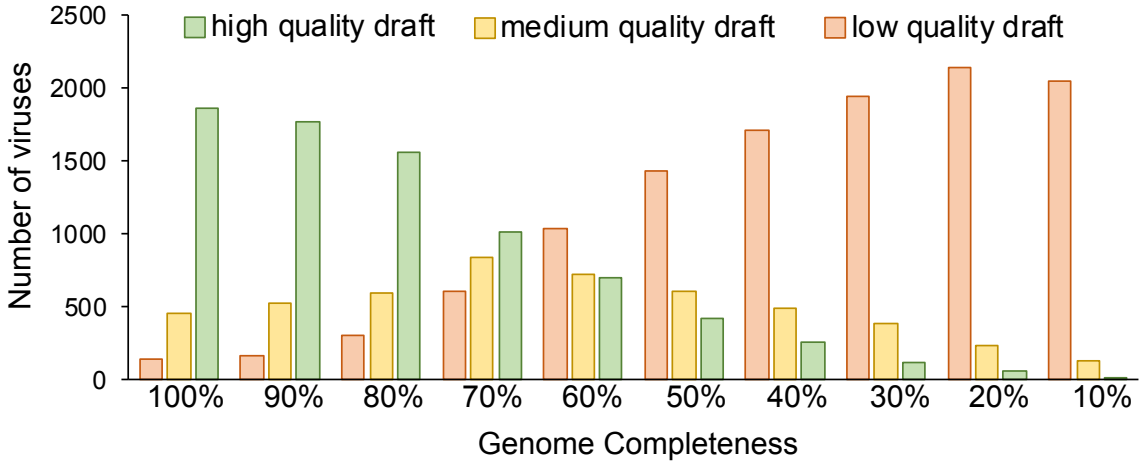
VIBRANT estimates the quality of predicted viral sequences, a relative proxy for completeness, and indicates sequences that are circular. To do this, VIBRANT uses annotation metrics of nucleotide replication and viral hallmark proteins. Hallmark proteins are those typically specific to viruses and those that are required for productive infection, such as structural (e.g., capsid, tail, baseplate), terminase or viral holin/lysin proteins. Nucleotide replication proteins are a variety of proteins associated with either replication or metabolism, such as nucleases, polymerases and DNA/RNA binding proteins. Viruses are categorized as low, medium or high-quality draft as determined by VOG annotations (Figure 5A, Additional File 17: Table S16). High-quality draft represents sequences that are likely to contain the majority of a virus's complete genome and will contain annotations that are likely to aid in analysis of the virus, such as phylogenetic relationships and true positive verification. Medium draft quality represents the majority of a complete viral genome but is more likely to be a smaller portion in comparison to high quality. These sequences may contain annotations useful for analysis but are under less strict requirements compared to high quality. Finally, low draft quality constitutes sequences that were not found to be of high or medium quality. Many metagenomic scaffolds will likely be low quality

genome fragments, but this quality category may still contain the higher quality genomes of some highly divergent viruses.

A

Quality	Interpretation
Complete circular	likely to be a complete viral genome
High quality draft	likey will represent the majority of a viral genome and contain information useful for analysis
Medium quality draft	likely will represent a significant portion of a viral genome and may contain information useful for analysis
Low quality draft	likely only a fragment of a viral genome

B



C

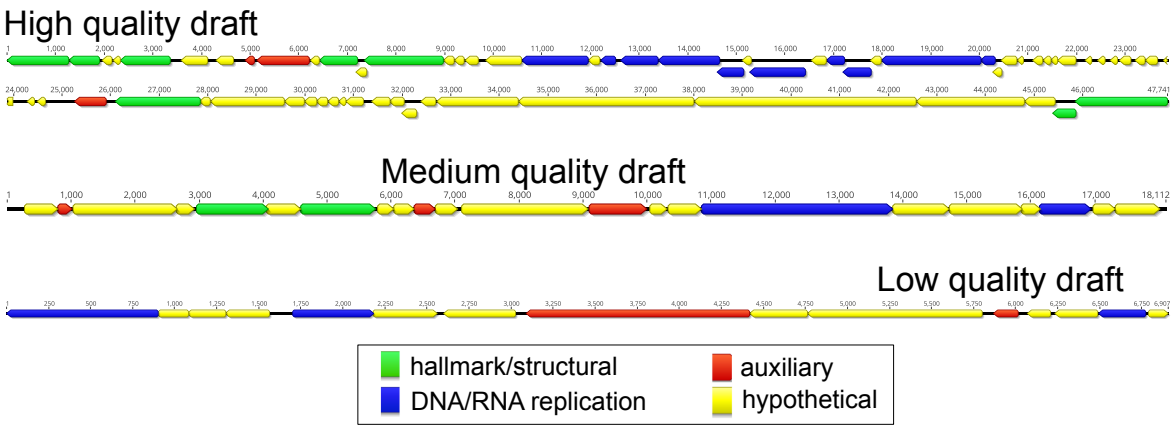


Figure 5. Estimation of genome quality of identified viral scaffolds. (A) Explanation of interpretation of quality categories: complete circular, high-quality draft, medium-quality draft and low-quality draft. Quality generally represents total proteins, viral annotations, viral hallmark protein and nucleotide replication proteins, which are common metrics used for manual verification of viral genomes. (B) Application of quality metrics to 2466 NCBI RefSeq *Caudovirales* viruses with decreasing genome completeness from 100% to 10% completeness, respective of total sequence length. All 2466 viruses are represented within each completeness group. (C) Examples of viral scaffolds representing low, medium and high-quality draft categories.

We benchmarked VIBRANT's viral genome quality estimation using a total of 2466 *Caudovirales* genomes from NCBI RefSeq database. Genomes were evaluated either as complete sequences or by removing 10% of the sequence at a time stepwise between 100% and 10% completeness (Figure 5B). The results of VIBRANT's quality analysis displayed a linear trend in indicating more complete genomes as high quality and less complete genomes as lower quality. The transition from categorizing genomes as high quality to medium quality ranged from 60% and 70% completeness. Although we acknowledge that VIBRANT's metrics are not perfect, we demonstrate the first benchmarked approach to quantify and characterize genome quality associated with completeness of viral sequences. Manual inspection and visual verification of viral genomes that were characterized into each of these genome quality categories showed that quality estimations matched annotations (Figure 5C).

Identifying function in viral communities: metabolic analysis

Viruses are a dynamic and key facet in the metabolic networks of microbial communities and can reprogram the landscape of host and community metabolism during infection. This can often be achieved by modulating host metabolic networks through expression of AMGs encoded on viral genomes. Identifying these AMGs and their associated role in the function of communities is imperative for understanding complex microbiome dynamics, or in some cases can be used to predict virus-host relationships. VIBRANT is optimized for the evaluation of viral community function by identifying and classifying the metabolic capabilities encoded by a virome. To do this, VIBRANT identifies AMGs and assigns them into specific metabolic pathways and broader categories as designated by KEGG annotations.

To highlight the utility of this feature we compared the metabolic function of IMG/VR viruses derived from several diverse environments: freshwater, marine, soil, human-associated and city (Additional File 16: Figure S2). We found natural environments (freshwater, marine and soil) to display a different pattern of metabolic capabilities compared to human environments (human-associated and city). Viruses originating from natural environments tend to largely encode AMGs for amino acid and cofactor/vitamin metabolism with a more secondary focus on carbohydrate and glycan metabolism. On the other hand, AMGs from city and human environments are dominated by amino acid metabolism, and to some extent cofactor/vitamin and sulfur relay metabolism. In addition to this broad distinction, all five environments appear slightly different from each other. Despite freshwater and marine environments appearing similar in the ratio of AMGs by metabolic category, the overlap in specific AMGs is less extensive. The dissimilarity between natural and human environments is likewise corroborated by the relatively low overlap in individual AMGs.

A useful observation provided by VIBRANT's metabolic analysis is that there appears to be globally conserved AMGs (i.e., present within at least 10 of the 12 environments tested). These 14 genes—*dcm, cysH, folE, phnP, ubiG, ubiE, waaF, moeB, ahbD, cobS, mec, queE, queD, queC*—likely perform functions that are central to viral replication regardless of host or environment. Notably, *folE*, *queD*, *queE* and *queC* constitute the entire 7-cyano-7-deazaguanine ($preQ_0$) biosynthesis pathway, but the remainder of queuosine biosynthesis are entirely absent with the exception of *queF*. Certain AMGs are unique in that they are the only common representatives of a pathway amongst all AMGs identified, such as *phnP* for methylphosphonate degradation. These AMGs may indicate an evolutionary advantage for manipulating a specific step of a pathway, such as overcoming a reaction bottleneck, as opposed to modulating an entire pathway

as seen with $preQ_0$ biosynthesis. However, it should be noted that this list of 14 globally conserved AMGs may not be entirely inclusive of the core set of AMGs in a given environment.

VIBRANT was evaluated for its ability to provide new insights into viral community function by highlighting AMGs from mixed metagenomes. Using only data from VIBRANT's direct outputs, we compared the viral metabolic profiles of 6 hydrothermal vent and 15 human gut metagenomes (Figure 6). As anticipated, based on IMG/VR environment comparisons, the metabolic capabilities between the two environments were different even though the number of unique AMGs was relatively equal (138 for hydrothermal vents and 151 for human gut). The pattern displayed by metabolic categories for each metagenome was similar to that displayed by marine and human viromes. For hydrothermal vents the dominant AMGs were part of carbohydrate, amino acid and cofactor/vitamin metabolism, whereas human gut AMGs

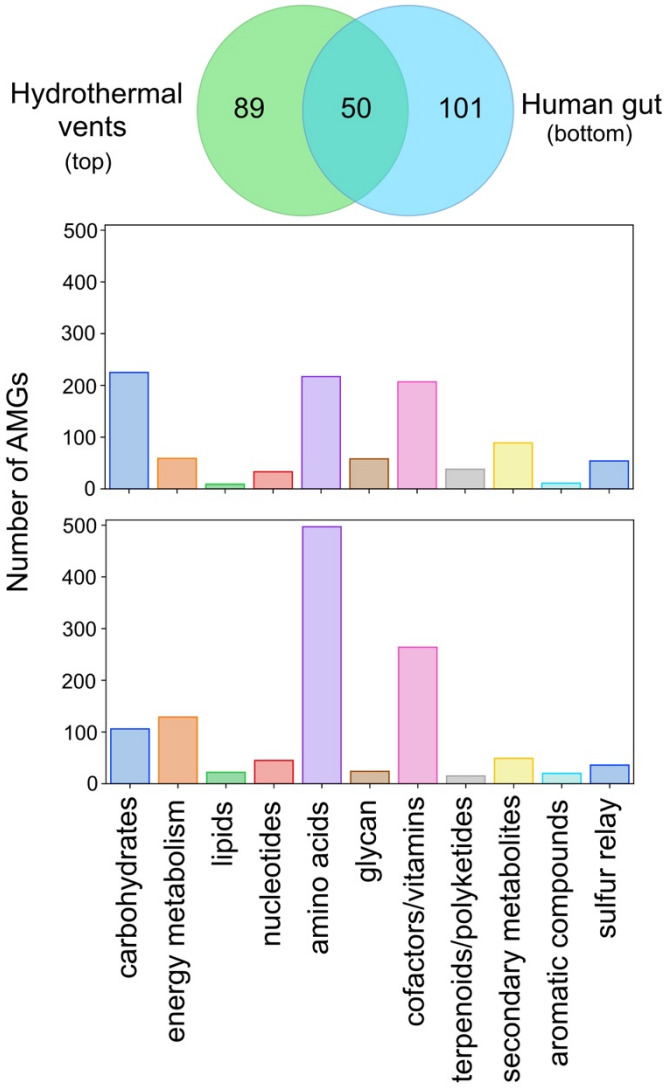


Figure 6. Comparison of AMG metabolic categories between hydrothermal vents and human gut. Venn diagram depicts the unique and shared non-redundant AMGs between 6 hydrothermal vent and 15 human gut metagenomes. The graphs depict the differential abundance of KEGG metabolic categories of respective AMGs for hydrothermal vents (top) and human gut (bottom).

were mostly components of amino acid and, to some extent, cofactor/vitamin metabolism. Although the observed AMGs and metabolic pathways were overall different, about a third (50 total AMGs) of all AMGs from each environment were shared; between these metagenomes alone all 14 globally conserved AMGs were present.

Observations of individual AMGs provided insights into how viruses interact within different environments. For example, tryptophan 7-halogenase (*prnA*) was identified in high abundance (45 total AMGs) within hydrothermal vent metagenomes but was absent from the human gut. Verification using GOV2 (Global Ocean Viromes 2.0) [312] and Human Gut Virome databases supported our finding that *prnA* appears to be constrained to aquatic environments, which is further supported by the gene's presence on several marine cyanophages. PrnA catalyzes the initial reaction for the formation of pyrrolnitrin, a strong antifungal antibiotic. Identification of this AMG only within aquatic environments suggests a directed role in aquatic virus lifestyles. Similarly, cysteine desulfhydrase (*iscS*) was abundant (14 total AMGs) within the human gut metagenomes but not hydrothermal vents.

Application of VIBRANT: Identification of viruses from individuals with Crohn's Disease

We applied VIBRANT to identify viruses of at least 5kb in length from 102 human gut metagenomes (discovery dataset): 49 from individuals with Crohn's Disease and 53 from healthy individuals [160,300]. VIBRANT identified 14,121 viruses out of 511,977 total scaffolds. These viral sequences were dereplicated to 8,822 non-redundant viral sequences using a cutoff of 95% nucleotide identity over at least 70% of the sequence. We next used read coverage of each virus sequence from all 102 metagenomes to calculate relative differential abundance across Crohn's Disease and healthy individuals. In total, we found 721 viral sequences to be more abundant in the

gut microbiomes associated with Crohn's Disease (Crohn's-associated) and 950 to be more abundant in healthy individuals (healthy-associated).

Using these viruses identified by VIBRANT we sought to identify taxonomic or host-association relationships to differentiate the viral communities of individuals with Crohn's Disease. We used vConTACT2 to cluster the 721 Crohn's- or 950 healthy-associated virus sequences with reference genomes using protein similarity. The majority of virus sequences (95.5%) were not clustered with any reference genome at approximately the genus level suggesting VIBRANT may have identified a large pool of novel or unique viral genomes. Although fewer total viruses were associated with Crohn's Disease, significantly more were clustered to at least one representative at the genus level (72 for Crohn's and 4 for healthy). Interestingly, no differentially abundant viruses from healthy individuals clustered with Enterobacterales-infecting reference viruses (enteroviruses), yet the majority (60/76) of Crohn's-associated viruses were clustered with known enteroviruses, such as Lambda- and Shigella-related viruses. The remaining 16 viruses mainly clustered with *Caudovirales* infecting *Lactococcus*, *Clostridium*, *Riemerella*, *Klebsiella* and *Salmonella* species, though *Microviridae* and a likely complete crAssphage were also identified. A significant proportion of all Crohn's-associated viruses (250/721), and the majority of genus-level clustered viruses (42/76), were found to be integrated sequences within a microbial genomic scaffold but were able to be identified due to VIBRANT's ability to excise proviruses.

We also generated a protein sharing network containing all 721 Crohn's and 950 healthy-associated virus sequences, which corresponded to taxonomic and host relatedness (Figure 7A). This protein network identified two different clustering patterns: (1) overlapping Crohn's and healthy-associated viral populations clustered with Firmicutes-like viruses which may be

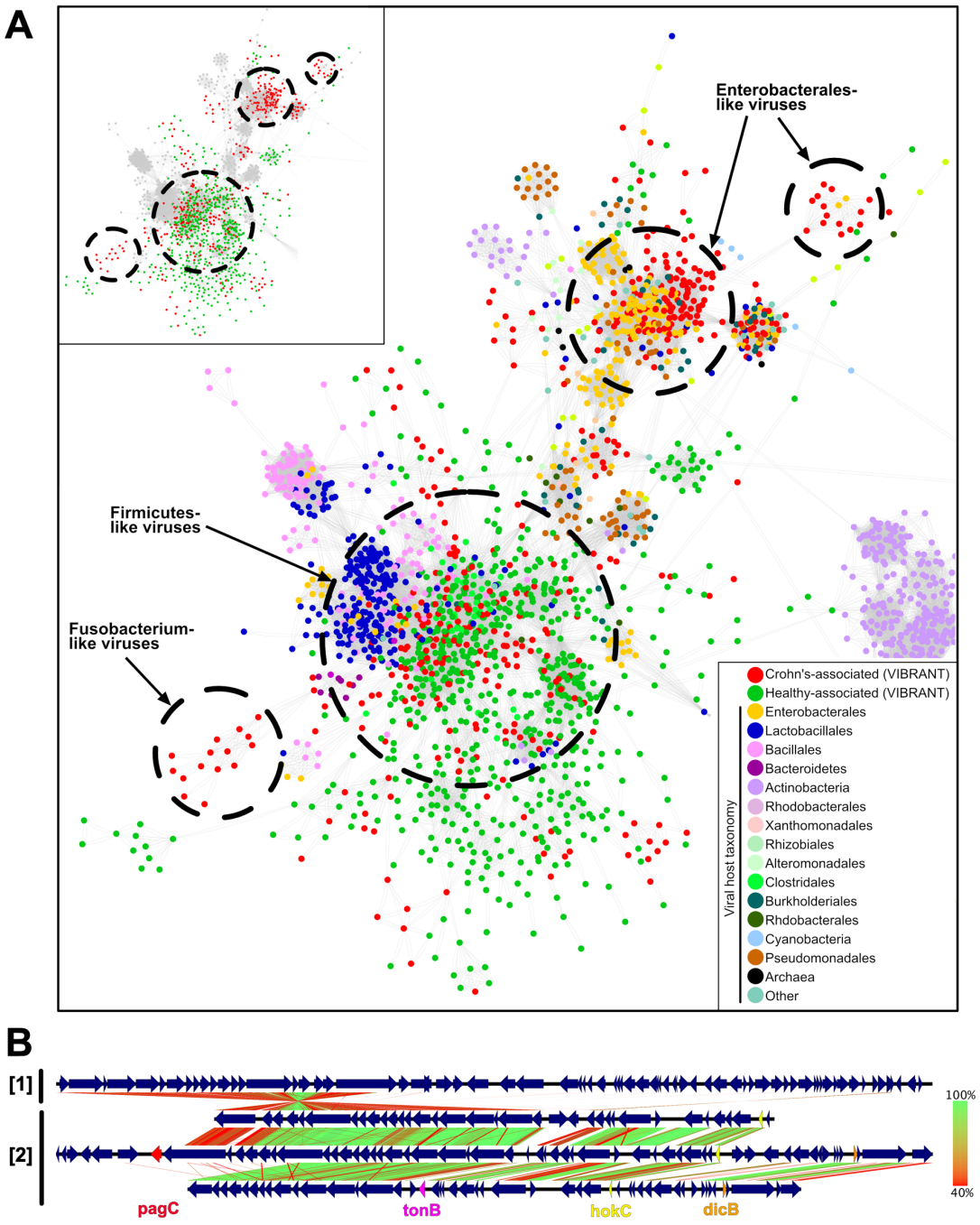


Figure 7. Viral metabolic comparison between Crohn's Disease and healthy individuals gut metagenomes. (A) Partial view of vConTACT2 protein network clustering of viruses identified by VIBRANT and reference viruses. Small clusters and clusters with no VIBRANT representatives are not shown. Each dot represents one genome and is colored according to host or dataset association. Relevant viral groups are indicated by dotted circles (circles enclose estimated boundaries). (B) tBLASTx similarity comparison between (1) Escherichia phage Lambda and (2) three Crohn's-associated viruses identified by VIBRANT. Putative virulence genes are indicated: *pagC*, *tonB*, *hokC* and *dicB*.

indicative of a stable gut viral community; (2) Crohn's-associated viruses clustered with Enterobacterales-like and Fusobacterium-like viruses which may be indicative of a state of dysbiosis. The presence of a greater diversity and abundance of Enterobacterales and Fusobacteria has previously been linked to Crohn's Disease [313,314], and therefore the presence of viruses infecting these bacteria may provide similar information.

VIBRANT provides annotation information for all of the identified viruses which can be used to infer functional characteristics in conjunction with host association. Comparison of Crohn's-associated Lambda-like virus genomic content and arrangement suggested a possible role of virally encoded host-persistence and virulence genes that are absent in the healthy-associated virome (Figure 7B). Among all Crohn's-associated viruses, 17 total genes (*bor, dicB, dicC, hokC, kilR, pagC, ydaS, ydaT, yfdN, yfdP, yfdQ, yfdR, yfdS, yfdT, ymfL, ymfM* and *tonB*) that have the potential to impact host survival or virulence were identified. Importantly, no healthy-associated viruses encoded such genes (Table 2). The presence of these putative dysbiosis-associated genes (DAGs) may contribute to the manifestation and/or persistence of disease, similar to what has been proposed for the bacterial microbiome [315–317]. For example, *pagC* encodes an outer membrane virulence factor associated with enhanced survival of the host bacterium within the gut [318]. The identification of *dicB* encoded on a putative *Escherichia* virus is unique in that it may represent a 'cryptic' provirus that protects the host from lytic viral infection, thus likely to enhance the ability of the host to survive within the gut [319]. Finally, *hokC* may indicate mechanisms of virally encoded virulence [320].

To characterize the distribution and association of DAGs with Crohn's Disease, we calculated differential abundance for two highly abundant DAG-encoding viruses across all metagenome samples. The first virus encoded *pagC* and *yfdN*, and the second encoded *dicB*, *dicC*

and *hokC*. Comparison of Crohn's Disease to healthy metagenomes indicates these viruses are present within the gut metagenomes of multiple individuals but more abundant in association with Crohn's Disease (Additional File 16: Figure S3A). This suggests an association of disease with not only putative DAGs, but also specific, and potentially persistent, viral groups that encode them. In order to correlate increased abundance with biological activity we calculated the index of replication (iRep) for each of the two viruses [156]. Briefly, iRep is a function of differential read coverage which is able to provide an estimate of active genome replication. Seven metagenomes containing the greatest abundance for each virus were selected for iRep analysis and indicated that each virus was likely active at the time of collection (Additional File 16: Figure S3B).

To validate these aforementioned findings, we applied VIBRANT to two additional metagenomic datasets from cohorts of individuals with Crohn's disease and healthy individuals (validation dataset): 43 from individuals with Crohn's Disease and 21 from healthy individuals [303,304]. VIBRANT identified 3,759 redundant viral genomes from Crohn's-associated metagenomes and 1,444 from healthy-associated metagenomes. Determination of protein networks and visualization similarly identified clustering of Crohn's-associated viruses with reference enteroviruses (Additional File 16: Figure S4). Likewise, we were able to identify 15 out of the 17 putative DAGs to be present in higher abundance in the Crohn's Disease microbiome (Additional File 18: Table S17). This validates our findings of the presence of unique viruses and proteins associated with Crohn's Disease, and suggests Enterobacterales-like viruses and putative DAGs may act as markers of Crohn's Disease. Overall, our results suggest that VIBRANT provides a platform for characterizing these relationships.

Discussion

Viruses that infect bacteria and archaea are key components in the structure, dynamics, and interactions of microbial communities [5,63,136,278,312]. Tools that are capable of efficient recovery of these viral genomes from mixed metagenomic samples are likely to be fundamental to the growing applications of metagenomic sequencing and analyses. Importantly, such tools would need to reduce bias associated with specific viral groups (e.g., *Caudovirales*) and highly represented environments (e.g., marine). Moreover, viruses that exist as integrated proviruses within host genomes should not be ignored as they can represent a substantial fraction of infections in certain conditions and also persistent infections within a community [311].

Here we have presented VIBRANT, a newly described method for the automated recovery of both free and integrated viral genomes from metagenomes that hybridizes neural network machine learning and protein signatures. VIBRANT utilizes metrics of non-reference based protein similarity annotation from KEGG, Pfam and VOG databases in conjunction with a unique 'v-score' metric to recover viruses with little to no biases. VIBRANT was built with the consideration of the human guided intuition used to manually inspect metagenomic scaffolds for viral genomes and packages these ideas into an automated software. This platform originates from the notion that proteins generally considered as non-viral, such as ribosomal proteins [72], may be decidedly common amongst viruses and should be considered accordingly when viewing annotations. V-scores are meant to provide a quantitative metric for the level of virus-association for each annotation used by VIBRANT, especially for Pfam and KEGG HMMs. That is, v-scores provide a means for both highlighting common or hallmark viral proteins as well as differentiating viral from non-viral annotations. In addition, v-scores give a quantifiable value to viral hallmark genes instead of categorizing them in a binary fashion.

VIBRANT was not only built for the recovery of viral genomes, but also to act as a platform for investigating the function of a viral community. VIBRANT supports the analysis of viromes by assembling useful annotation data and categorizing the metabolic pathways of viral AMGs. Using annotation signatures, VIBRANT furthermore is capable of estimating genome quality and distinguishing between lytic and lysogenic viruses. To our knowledge, VIBRANT is the first software that integrates virus identification, annotation and estimation of genome completeness into a stand-alone program.

Benchmarking and validation of VIBRANT indicated improved performance compared to VirSorter [115], VirFinder [113] and MARVEL [293], three commonly used programs for identifying viruses from metagenomes. This included a substantial increase in the relationship between true virus identifications (recall, true positive rate) and false non-virus identifications (specificity, false positive rate). That is, VIBRANT recovered more viruses with no discernable expense to false identifications. The result was that VIBRANT was able to recover an average of 2.3 and 1.7 more viral sequence from real metagenomes than VirFinder and VirSorter, respectively. When tested on metagenome-assembled viral genomes from IMG/VR [262] representing diverse environments VIBRANT was found to have no perceivable environment bias towards identifying viruses. In comparison to provirus prediction tools, specifically PHASTER [295], Prophage Hunter [296] and VirSorter, VIBRANT was shown to be proficient in identifying viral regions within bacterial genomes. This included the identification of a putative *Bacteroides* provirus that PHASTER and Prophage Hunter were unable to identify. The importance of integrated provirus prediction was underscored in the analysis of Crohn's Disease metagenomes since it was found that a significant proportion of disease related viruses were temperate viruses existing as host-integrated genomes.

VIBRANT's method allows for the distinction between scaffold size and coding capacity in designating the minimum length of virus identifications. Traditionally, a cutoff of 5000 bp has been used to filter for scaffolds of a sufficient length for analysis. This is under the presumption that a longer sequence will be likely to encode more proteins. For example, this cutoff has been adopted by IMG/VR. However, we suggest a total protein cutoff of four open reading frames rather than sequence length cutoff to be more suitable for comprehensive characterization of the viral community. VIBRANT's method works as a strict function of total encoded proteins and is completely agnostic to sequence length for analysis. Therefore, the boundary of minimum encoded proteins will support a more guided cutoff for quality control of virus identifications. For example, increasing the minimum sequence length to 5000 bp will have no effect on accuracy or ability to recall viruses since VIBRANT will only be considerate of the minimum total proteins, which is set to four. The result will be the loss of all 1000 bp to 4999 bp viruses that still encode at least four proteins. To visualize this distinction, we applied VIBRANT with various length cutoffs to the previously used estuary virome (see Table 1). Input sequences were stepwise limited from 1000 bp to 10000 bp (1000 bp steps) or four open reading frames to 13 open reading frames (one open reading frame steps) in length. Limiting to open reading frames indicated a reduced drop-off in total virus identifications and total viral sequence compared to a minimum sequence length limit (Additional File 16: Figure S5).

The output data generated by VIBRANT—protein/gene annotation information, protein/gene sequences, HMM scores and e-values, viral sequences in FASTA and GenBank format, indication of AMGs, genome quality, etc.—provides a platform for easily replicated pipeline analyses. Application of VIBRANT to characterize the function of Crohn's-associated viruses emphasizes this utility. VIBRANT was not only able to identify a substantial number of

viral genomes, but also provided meaningful information regarding putative DAGs, viral sequences for differential abundance calculation and genome alignment, viral proteins for clustering, and AMGs for metabolic comparisons.

Conclusions

Our construction of the VIBRANT platform expands the current potential for virus identification and characterization from metagenomic and genomic sequences. When compared to two widely used software programs, VirFinder and VirSorter, we show that VIBRANT improves total viral sequence and protein recovery from diverse human and natural environments. As sequencing technologies improve and metagenomic datasets contain longer sequences VIBRANT will continue to outcompete programs built for short scaffolds (e.g., 500-3000 bp) by identifying more higher quality genomes. Our workflow, through the annotation of viral genomes, aids in the capacity to discover how viruses of bacteria and archaea may shape an environment, such as driving specific metabolism during infection or dysbiosis in the human gut. Furthermore, VIBRANT is the first virus identification software to incorporate annotation information into the curation of predictions, estimation of genome quality and infection mechanism (i.e., lytic vs lysogenic). We anticipate that the incorporation of VIBRANT into microbiome analyses will provide easy interpretation of viral data, enabled by VIBRANT's comprehensive functional analysis platform and visualization of information.

Table 1. Virus recovery of VIBRANT, VirFinder and VirSorter from mixed metagenomes and a virome. Mixed community assembled metagenomes from human gut, thermophilic compost and freshwater, as well as an estuary virome, were used to compare virus prediction ability between the three programs. For each assembly the scaffolds were limited to a minimum length of 1000bp. Only a subset of each dataset contained scaffolds encoding at least four open reading frames. VIBRANT, VirFinder (score minimum of 0.90) and VirSorter (categories 1 and 2) were compared by total viral predictions, total combined length of predicted viruses, and total combined proteins of predicted viruses. Comparison columns, denoted "VIBRANT vs. VirFinder" and "VIBRANT vs. VirSorter", display the comparison ratio of the given metric; bold indicates greater performance by VIBRANT.

Metagenome	seqs. total (≥1kb)	Seqs. ≥ 4 ORFs	Metric	VIBRANT	VirFinder (score≥0.90)	VIBRANT vs. VirFinder	VirSorter (cat. 1 & 2)	VIBRANT vs. VirSorter
human gut: adenoma	34,883	11,360	total putative viruses	527	604	0.87	284	**1.86**
			total virus length (bp)	5,234,242	1,696,118	**3.09**	3,982,292	**1.31**
			total virus proteins	7,661	2,134	**3.59**	5,484	**1.40**
human gut: carcinoma	53,946	18,669	total putative viruses	784	1,329	0.59	450	**1.74**
			total virus length (bp)	5,611,953	3,500,838	**1.60**	4,182,862	**1.34**
			total virus proteins	8,401	4,644	**1.81**	5,945	**1.41**
human gut: healthy	42,739	17,079	total putative viruses	565	672	0.84	309	**1.83**
			total virus length (bp)	5,623,082	2,411,049	**2.33**	4,512,571	**1.25**
			total virus proteins	8,202	3,230	**2.54**	6,127	**1.34**
thermophilic compost	68,815	21,620	total putative viruses	1,047	878	**1.19**	383	**2.73**
			total virus length (bp)	10,253,162	2,238,129	**4.58**	3,290,654	**3.12**
			total virus proteins	9,912	2,806	**3.53**	4,400	**2.25**
freshwater lake (bog)	79,862	26,832	total putative viruses	5,626	7,567	0.74	1,503	**3.74**
			total virus length (bp)	34,976,570	25,357,664	**1.38**	15,436,797	**2.27**
			total virus proteins	56,120	37,537	**1.50**	21,280	**2.64**
*** estuary virome**	5,247	3,277	total putative viruses	3,141	2,294	**1.37**	1,121	**2.80**
			total virus length (bp)	6,591,285	6,478,804	**1.02**	5,163,674	**1.28**
			total virus proteins	20,500	12,035	**1.70**	9,645	**2.13**

* VIBRANT, VirFinder and VirSorter ran with alternate settings

Table 2. Identification of putative DAGs encoded by Crohn's-associated viruses. The differential abundance between Crohn's Disease and healthy metagenomes of 17 putative DAGs. Abundance of each gene represents non-redundant annotations, or total gene copy number, from Crohn's-associated and healthy-associated viruses.

ID	Gene	Name	Crohn's Disease	Healthy
PF06291.11	*bor*	Bor protein	8	0
K22304	*dicB*	cell division inhibition protein	8	0
K22302	*dicC*	transcriptional repressor of cell division inhibition gene dicB	18	0
K18919	*hokC*	protein HokC/D	16	0
VOG11478	*kilR*	Killing protein	15	0
K07804	*pagC*	putatice virulence related protein	13	0
PF15943.5	*ydaS*	Putative antitoxin of bacterial toxin-antitoxin system	22	0
PF06254.11	*ydaT*	Putative bacterial toxin	18	0
VOG04806	*yfdN*	Uncharacterized protein	19	0
VOG01357	*yfdP*	Uncharacterized protein	11	0
VOG11472	*yfdQ*	Uncharacterized protein	11	0
VOG01639	*yfdR*	Uncharacterized protein	17	0
VOG01103	*yfdS*	Uncharacterized protein	18	0
VOG16442	*yfdT*	Uncharacterized protein	8	0
VOG00672	*ymfL*	Uncharacterized protein	25	0
VOG21507	*ymfM*	Uncharacterized protein	9	0
K03832	*tonB*	periplasmic protein	3	0

Data Availability

VIBRANT is implemented in Python and all scripts and associated files are freely available at https://github.com/AnantharamanLab/VIBRANT/. The datasets supporting the conclusions of this article are included within the article and its additional files (Additional File 1: Table S1 and Additional File 20: Table S19). VIBRANT is also freely available for use as an application through the CyVerse Discovery Environment; to use the application visit https://de.cyverse.org/de/. Additional details of relevant data are available from the corresponding author on request.

Acknowledgements

We thank Upendra Devisetty for his assistance with dockerizing and integrating VIBRANT as a web-based application in the CyVerse Discovery Environment. We thank the University of Wisconsin - Office of the Vice Chancellor for Research and Graduate Education, University of

Wisconsin – Department of Bacteriology, and University of Wisconsin – College of Agriculture and Life Sciences for their support.

Author Contributions

K.K and K.A designed the study, performed all analyses and interpretation of data, and wrote the manuscript. Z.Z contributed to conceptualization of study design and reviewed the manuscript. All authors have reviewed and approved the final manuscript.

Chapter 5: Deciphering active prophages from metagenomes

Kristopher Kieft[1,2] and Karthik Anantharaman[1]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA
[2]Microbiology Doctoral Training Program, University of Wisconsin–Madison, Madison, WI, USA

Publication:
Kieft, K. & Anantharaman, K. Deciphering Active Prophages from Metagenomes. *mSystems* **0**, e00084-22 (2022).

All supplementary figures, tables, and files are available at the following Figshare repository:
https://figshare.com/projects/Kristopher_Kieft_PhD_Dissertation/136427

Abstract

Temperate phages (prophages) are ubiquitous in nature and persist as dormant components of host cells (lysogenic stage) before activating and lysing the host (lytic stage). Actively replicating prophages contribute to central community processes, such as enabling bacterial virulence, manipulating biogeochemical cycling, and driving microbial community diversification. Recent advances in sequencing technology have allowed for the identification and characterization of diverse phages, yet no approaches currently exist for identifying if a prophage has activated. Here, we present PropagAtE (Prophage Activity Estimator), an automated software tool for estimating if a prophage is in the lytic or lysogenic stage of infection. PropagAtE uses statistical analyses of prophage-to-host read coverage ratios to decipher actively replicating prophages, irrespective of whether prophages were induced or spontaneously activated. We demonstrate that PropagAtE is fast, accurate and sensitive, regardless of sequencing depth. Application of PropagAtE to prophages from 348 complex metagenomes from human gut, murine gut and soil environments identified distinct spatial and temporal prophage activation signatures, with the highest proportion of active prophages in murine gut samples. In infants treated with antibiotics or infants without treatment, we identified active prophage populations correlated to specific treatment groups. Within time series samples from the human gut, 11 prophage populations, some encoding the sulfur metabolism gene *cysH* or a *rhuM*-like virulence factor, were consistently present over time but not active. Overall, PropagAtE will facilitate accurate representations of viruses in microbiomes by associating prophages with their active roles in shaping microbial communities in nature.

Importance

Viruses that infect bacteria are key components of microbiomes and ecosystems. They can kill and manipulate microorganisms, drive planetary-scale processes, biogeochemical cycling, and influence the structures of entire food networks. Prophages are viruses that can exist in a dormant state within the genome of their host (lysogenic stage) before activating in order to replicate and kill the host (lytic stage). Recent advances have allowed for the identification of diverse viruses in nature, but no approaches exist for characterizing prophages and their stages of infection (prophage activity). We develop and benchmark an automated approach, PropagAtE (Prophage Activity Estimator), to identify the stages of infection of prophages from genomic data. We provide evidence that active prophages vary in identity and abundance across multiple environments and scales. Our approach will enable accurate and unbiased analyses of viruses in microbiomes and ecosystems.

Introduction

Viruses that infect bacteria and archaea (bacteriophages or phages) are pervasive entities that are ubiquitous on Earth. Phages drive evolutionary adaptation and diversification of microorganisms, play critical roles in global nutrient cycles and can directly impact human health [5,21,48,136,137,147,277,278]. Phages can be organized into two categories according to how they infect a host cell: lytic and temperate. Temperate phages are those that have the ability to integrate their dsDNA genome into their bacterial host and can be identified in nearly half of all cultivated bacteria [321]. These integrated prophage sequences can coexist with the host cell in a lysogenic stage in which virions are not produced. During host genome replication the prophage sequence is likewise replicated in a one-to-one ratio. Given host-dependent or environmental cues such as DNA damage or nutrient stressors, or spontaneous activation, the prophage can enter a lytic stage

to produce virions and lyse the host [112,322–326]. On the other hand, lytic phages are those that directly enter the lytic stage upon infection with no mechanism for integration and dormancy.

Prophages can affect their host and surrounding microbial communities in both the "dormant" lysogenic stage as well as in the "active" lytic stage. In the dormant stage, prophages can impose physiological changes on the host by altering gene expression patterns, inducing DNA transfer or recombination events, and providing virulence attributes [39,327–330]. For example, the pathogenicity of some strains of *Staphylococcus aureus* is reliant on the presence of integrated prophage sequences [55]. In the active stage, the result of phage lysis significantly impacts microbial communities by turning over essential nutrients, especially carbon, nitrogen and sulfur [35,37,69,82,84,85]. Lysis of bacterial populations likewise alters whole microbiomes by diversifying community structures and expanding niche opportunities [136,257]. For example, the "Kill the Winner" model of virus population growth suggests that dominant bacterial populations are more susceptible to phage predation, which will facilitate expansions of less abundant taxa as the dominant populations are lysed [58,59,331]. Despite the importance of phage lysis on microbial communities, the proportion of lysis by prophages entering the lytic cycle is unclear. As opposed to strictly lytic phages, it remains difficult to associate prophages with active lysis. This is because prophage genome abundance can fluctuate according to host genome replication in the absence of lysis, whereas lytic phages, with few exceptions, must lyse a host in order to increase the abundance of their genomes.

In addition to traditional approaches such as isolation of phages, advances in high throughput metagenomic sequencing have sped up the ability to identify a large diversity of lytic and lysogenic phage sequences. Recently developed software have allowed for accurate characterization of prophages in both isolate and metagenomic assembled genomes, namely

VIBRANT [117], VirSorter [115], PHASTER [295] and Prophage Hunter [296]. Thus far these software have allowed us to begin to estimate the total diversity of prophages in nature. However, identifying the genome sequences of prophages does not provide context to their *in situ* state of being in the lysogenic or lytic stage of infection. This information is vital as it distinguishes which prophage or phage populations are actively impacting a microbial community through lysis events. Moreover, with the exception of Prophage Hunter, current software cannot distinguish prophage genomes that have become "cryptic", or those that have lost functional abilities to enter the lytic stage [319,332,333]. Yet, Prophage Hunter still cannot identify if a given prophage is active, only if it may have the ability to do so.

Providing context to the infection stage of a prophage is imperative for accurate conclusions on its role in effecting its host and the microbial community. For example, identifying a prophage encoding a virulence factor or metabolic gene may have important implications for its role in manipulating its host's pathogenic interactions, metabolic transformations, and impacts on nutrient and biogeochemical cycling. In order to place the prophage into context within the microbial community it would be necessary to first determine which stage the prophage is in, namely lytic or lysogenic. Assuming that all identified prophages are in a lytic stage could lead to misrepresentations or misinterpretations of the data if the prophage is actually dormant, or even cryptic.

Here, we present the software PropagAtE (Prophage Activity Estimator). PropagAtE uses genomic coordinates of integrated prophage sequences and short sequencing reads to estimate if a given prophage was in the lysogenic (dormant) or lytic (active) stage of infection. PropagAtE was designed for use with metagenomic data but can also use other forms of genomic data (e.g., sequence data from isolated microorganisms). When tested on systems with known active

prophages PropagAtE was fully accurate in determining prophages that were active versus dormant, regardless of read coverage depth. No active prophages were identified in control systems encoding prophages that were known to be dormant. PropagAtE was also utilized to identify active prophages in several metagenomes, including the adult and infant human gut, murine gut, and three different peatland soil environments. We show that specific prophages can be identified within differing antibiotic treatment and no treatment groups of individuals, and that activity of those prophages are correlated to particular treatment groups. Finally, we show that identifying the retention of a prophage over time does not necessarily indicate activity over time. PropagAtE is freely available at https://github.com/AnantharamanLab/PropagAtE.

Results

Conceptualization of PropagAtE

Temperate phages that are integrated exist as a component of their host's genome. When the host genome replicates, the prophage is also replicated likewise in a one-to-one ratio. As a result, when sequencing the host genome the prophage region and the flanking host region(s) are represented equally. Upon activation and entry into the lytic cycle, the prophage sequence is independently replicated for phage propagation and assembly into new virions. At this stage within the host cell there will be one host genome equivalent for multiple phage genomes regardless of whether lysis has occurred yet or not. Following lysis, virions containing phage genomes are released into the surrounding environment. These released genomes continue to represent the ratio of prophage to host genome copies if these prophage genomes are still included in the metagenome (**Fig. 1A**).

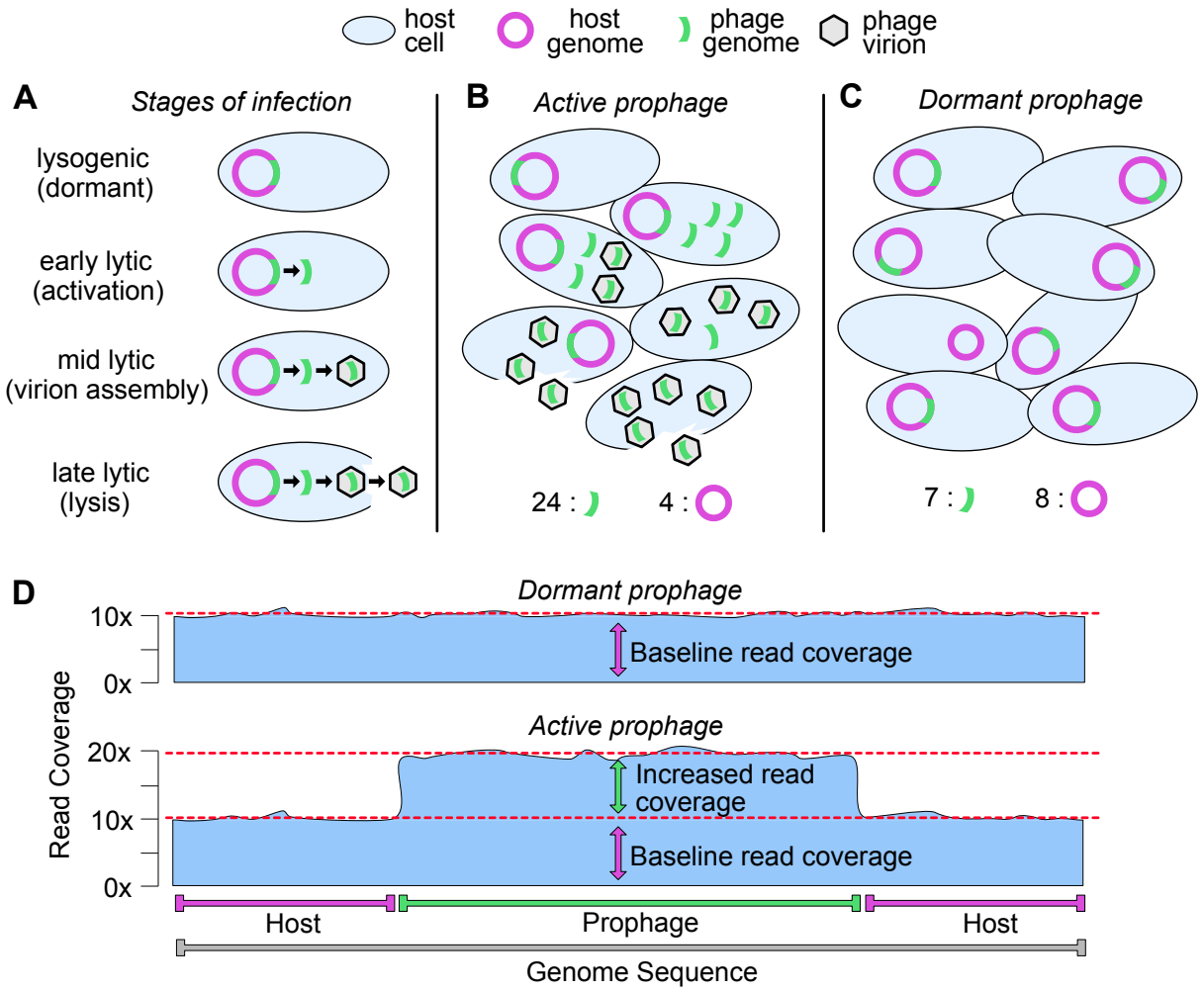


Figure 1. Schematic conceptualization of PropagAtE mechanism. (A) Stages of integrated prophage infection from the lysogenic (dormant) to lytic (active) stages. Over the course of infection the prophage:host genome copy ratio increases. (B) Microbial community structure with an active prophage, from phage activation to lysis. The prophage:host genome copy ratio increases to greater than 1:1 through phage genome replication and host genome degradation. (C) Microbial community structure with a dormant prophage in which the prophage:host genome copy ratio is near 1:1. Here, one host is depicted as having cured the prophage from its genome. (D) Conceptual diagram of the read coverage for a prophage in a dormant (top) or active (bottom) stage of infection. Active prophages result in an increased read coverage above the baseline read coverage of the host.

The specific ratio of phage to host genomes depends on many factors. One major factor is the burst size of a given phage, or the number of virions released from a lysed host. Phage burst sizes can range from fewer than ten in the case of crAssphage that infects *Bacteroides intestinalis* [334], to many thousands in the case of phage MS2 that infects *Escherichia coli* [335]. Another factor,

utilized by many phages including those that infect marine cyanobacteria, is that the host genome is degraded during the lytic stage to supply nucleotides to the replicating phage genomes which will further increase the prophage to host genome copy ratio [336,337]. Thus, during the lytic stage of phage propagation as well as post-lysis, the ratio of prophage to host genome copies will become skewed in favor of prophage genomes [338,339]. This will lead to a prophage:host genome copy ratio significantly greater than 1:1 (**Fig. 1B**). If the prophage was in a dormant stage of infection, the prophage:host genome copy ratio would be approximately 1:1 (**Fig. 1C**). This is likewise dependent on various factors, such as the ability of some members of the host population to "cure" (i.e., remove) the prophage from its genome. Despite nuances in specific prophage:host genome copy ratios, active prophages will yield a ratio greater than 1:1 whereas dormant prophages will yield a ratio near 1:1.

Whether or not the prophage:host genome copy ratio is skewed can be identified using statistical analyses of aligned sequencing read coverage after genome sequencing and read alignment. After sequencing and assembly of a system (e.g., isolated bacteria culture, complex microbiome, etc.), the integrated prophage sequence will assemble as a component of the host genome in a ~1:1 ratio, regardless of activity. However, if a prophage has activated then the resulting phage genome copies contained in virions are identical to the integrated prophage sequence. Therefore, read alignment to the assembly will recruit reads to the prophage and host regions in a ratio indicative of the stage of infection. During the lysogenic stage where the prophage is dormant, read recruitment will generate even coverage across the regions. Conversely, a prophage that has entered the lytic, active stage will generate an uneven read recruitment skewed towards greater coverage at the prophage region only (**Fig. 1D**). Read alignment will not determine

the true prophage:host abundance, but it can quantify a relative ratio to accurately determine stage of infection.

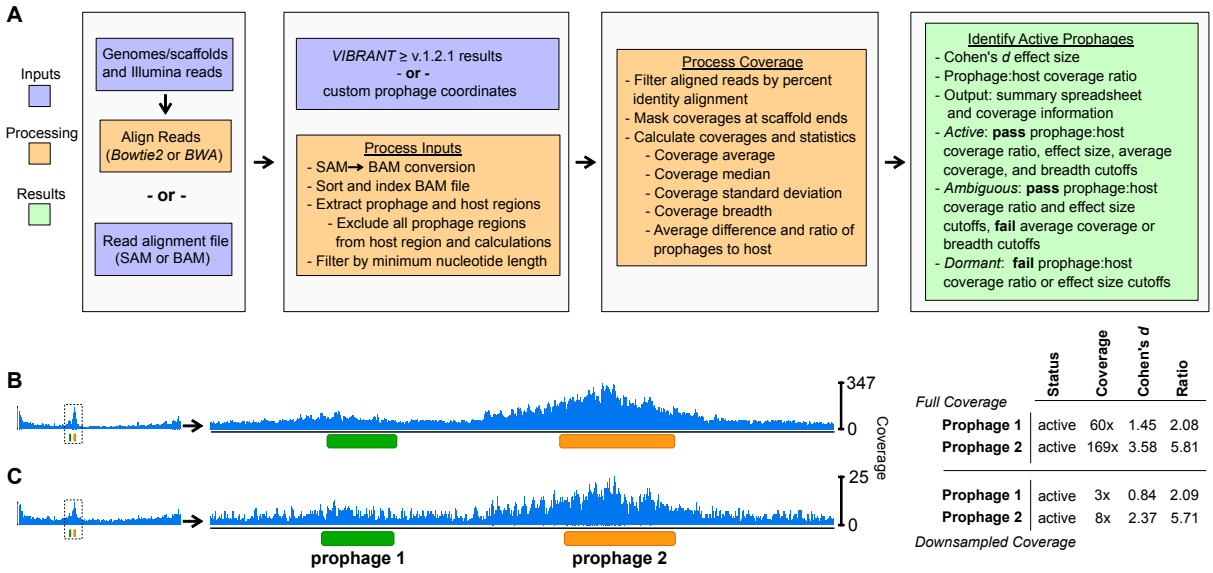


Figure 2. Workflow and implementation of PropagAtE. (A) Workflow of PropagAtE including data input, read alignment processing and results output. Example of read coverage profiles for two active *Bacillus licheniformis* DSM13 prophages with all reads (B) or 5% subsampled reads (C) aligned, respective to the conceptual diagram in **Figure 1D**. For B and C, statistics for coverage, Cohen's *d* effect size, and prophage:host coverage ratio are shown.

Overview of PropagAtE's workflow

Differentiating active prophages from those that are dormant is essential for accurate representation and evaluation of individual-cell and community-level systems. PropagAtE provides the first automated platform for the identification of active prophages that is scalable for isolate genomes or complex metagenomes. Since most prophages exist as an integrated (i.e., connected) element of a host genome, the read coverage from the prophage and host sections can be compared in a one-to-one manner to estimate a genome copy ratio. PropagAtE utilizes the ratio of prophage:host read coverage along with the ratio's effect size (i.e., significance of the ratio) to designate if a given prophage was dormant or active. The PropagAtE workflow can be simplified

into four general steps: data input, read alignment and processing, coverage calculations, and statistical results output (**Fig. 2A**). Users are given two options for data input: (1) genomes/scaffolds of host sequences with raw short sequencing reads or (2) a pre-generated alignment file in SAM or BAM format. If given the former input, reads will be aligned using Bowtie2 [225] to generate a SAM file. All SAM format files are converted to BAM format for more efficient processing [340].

Using the BAM file, either generated or supplied by the user, aligned reads exceeding the percent alignment threshold are removed. Following filtering, coverage per nucleotide is extracted including all nucleotides with zero coverage. To eliminate noise, coverage values at the sequence ends are trimmed off to a length roughly equivalent to the input read length. Then, users are given two options for prophage coordinate data input: direct results from a VIBRANT (v1.2.1 or greater) analysis [117] or a manually generated coordinate file of a specified format. In cases for which multiple prophages are present on a single genome/scaffold, all prophage regions are considered independently. In addition, the host region is segmented to exclude all prophage regions, but each segment is considered as a single, cohesive host sequence. That is, if two or more prophages are present on a single host scaffold, neither prophage will interfere with the other in terms of coverage value calculations and each prophage is compared to an identical prophage-excluded host region.

For each prophage and host pair, metrics for average coverages, median coverages, coverage standard deviations and prophage:host coverage ratio are calculated. Each prophage's activity is estimated according to the prophage:host coverage ratio and Cohen's *d* effect size of the coverage difference. Prophages exceeding the default or user-set thresholds for both metrics are considered as potentially active. Additionally, potentially active prophages must pass the minimum average coverage and minimum coverage breadth thresholds. If these latter coverage criteria are

met the prophage is estimated to be active, otherwise the prophage is labeled as ambiguous (**Fig. 1A**).

Read alignment can visualize active prophages

Two activated prophages in the genome of *Bacillus licheniformis* DSM13 [339] were used to visualize active prophage identification using PropagAtE using full and subsampled read sets (**Figs. 2B,C**). Visualization of the read coverage at each nucleotide in the genome clearly depicted coverage spikes exclusively at the prophage regions. The example prophages existed in close proximity to each other and had differing average coverages (60x and 169x). Both example prophages likewise met the minimum prophage:host coverage ratio (2.08 and 5.81) and Cohen's *d* effect size (1.45 and 3.58) thresholds. These results are in line with the conceptualization of the workflow seen in **Fig. 1D** apart from notable spikes in coverage at prophage genome centers and host genome ends. The host genome end coverage spikes are commonly explained by the location of the host's origin of replication [156,341]. The coverage spike at the prophage genome center is likely the result of a similar occurrence of a prophage replication-related packaging site [339,342].

Positive control tests for prophages from isolate genomes

Positive control tests were utilized in order to set threshold boundaries for PropagAtE to identify active prophages as well as assess the recall rate of PropagAtE. Positive control samples were considered as those for which DNA from both an active prophage and its host were extracted and sequenced in tandem. This method best represents metagenomic samples in which all DNA is extracted and sequenced together. In addition, extraction of both host and free phage DNA together is essential for positive tests because this method will best depict the most accurate prophage:host

coverage ratio. Three model systems for which sequencing data was publicly available were identified for use as positive controls. All experiments and sequencing were performed elsewhere [339,343,344] (**Supplemental Table S1**). Each system, since they represent isolate bacteria, have a much higher read coverage compared to a typical metagenome assembled genome. To ensure validation of PropagAtE for both isolate and metagenomic samples, two tests per system were done. One was done with all available reads ("full reads") and another was done with a random subset of 5% of the reads ("5% reads"). Furthermore, prophages were predicted from these systems using both VIBRANT and PHASTER to ensure accurate predictions despite variable prophage coordinate predictions. All PropagAtE results for positive control tests can be found in **Supplemental Table S2**.

The first system we tested was *Bartonella krasnovii* OE1-1 and its prophage [343]. In triplicate, the bacteria were either induced for prophage using mitomycin C or uninduced as controls. For the induced prophages, the prophage:host coverage ratios were relatively even between the three samples for VIBRANT (1.82, 1.87 and 2.07) and PHASTER (1.22, 1.26 and 1.21). Likewise, in the uninduced control samples the prophage:host coverage ratios depicted nearly equal coverage (VIBRANT: 1.06 and 1.13; PHASTER: 0.73, 0.98 and 1.03) except one sample from VIBRANT with a low ratio (0.46) (**Figs. 3A, B**). This suggests the method is reliable across multiple samples or time points for the same phage. The ratio effect size, using Cohen's *d* metric, indicated that the prophage:host coverage ratios observed from the VIBRANT predictions were significant in their difference. For the induced prophages the effect sizes were greater than one (1.20, 1.19, 1.15) indicating a high dissimilarity between the prophage and host coverages. The uninduced controls' effect sizes were low (0.33 and 0.59) except for the sample with the low ratio which had a higher effect size (1.62) corresponding to the host having a higher coverage (**Fig.**

3C). For PHASTER the same results were not observed. The effect sizes for both the induced prophages (0.45, 0.42 and 0.38) and uninduced controls (0.95, 0.07 and 0.15) were not significant (**Fig. 3D**). When 5% of the reads were randomly sampled for PropagAtE, the induced and uninduced results were essentially equivalent to that of the full read set for VIBRANT and PHASTER in terms of prophage:host coverage ratios (**Supplemental Figs. 1A, B**) and only marginally lower effect sizes (**Supplemental Figs. 1C, D**). This further indicates that high read coverage is not essential, nor significantly impacts, the outcome of analysis. However, this system suggests that the method in which prophages are predicted can determine the outcome and accuracy of PropagAtE activity estimation. Here, VIBRANT predictions yielded expected results whereas PHASTER predictions yielded dormant predictions where active was expected.

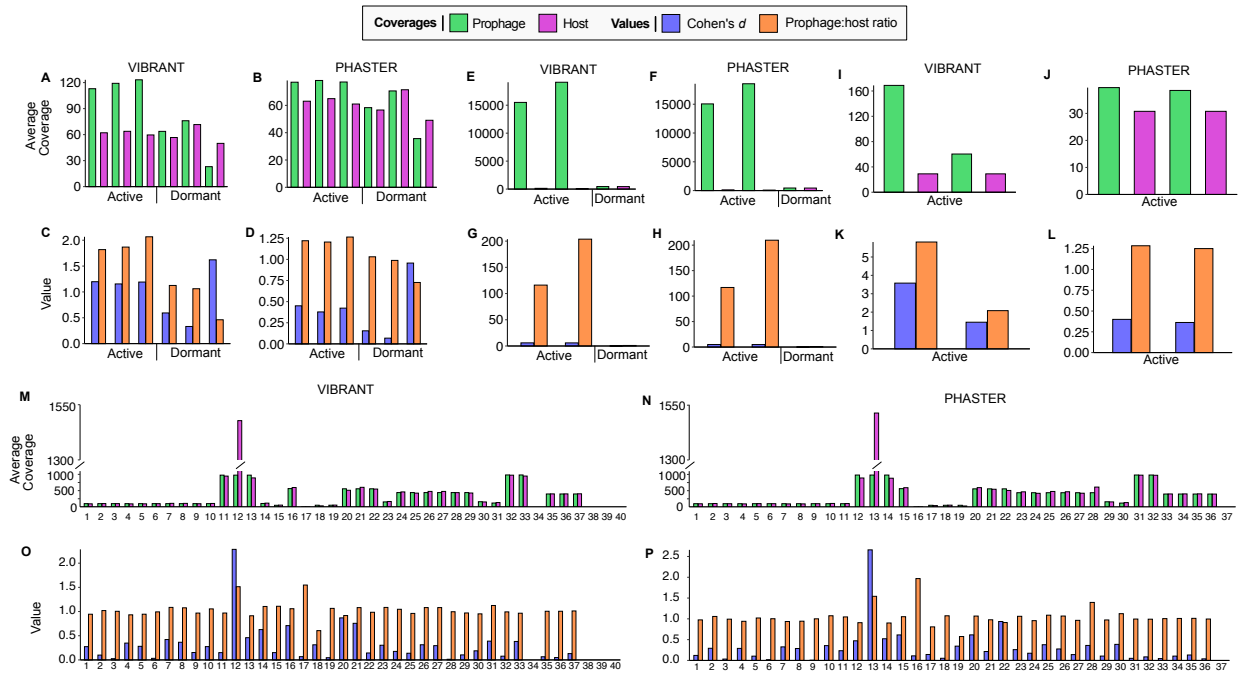


Figure 3. **Positive and negative control results using full read sets**. Positive control results for (A-D) *Bartonella krasnovii* OE1-1, (E-H) *Lactococcus lactis* MG1363, and (I-L) *Bacillus licheniformis* DSM13. Samples are labeled as containing active or dormant prophages. (M-P) All negative control results with each value on the x-axis representing a single prophage. Prophage and host average read coverages (green and purple, respectively) as well as Cohen's *d* effect sizes and prophage:host coverage ratios (blue and orange, respectively) are shown. Each positive and negative control set has prophage predictions generated by both VIBRANT and PHASTER (labeled vertically).

The second system we tested was *Lactococcus lactis* MG1363 and its prophage [344]. Similar to the previous system, in one sample the prophage was induced with mitomycin C and another was used as an uninduced control. The induction sample was sequenced 1- and 2-hours post-induction for a total of two positive samples. For the induced samples the resulting prophage:host coverage ratios were high and increased over time (VIBRANT: 116 and 204; PHASTER: 117 and 210). In the uninduced control the prophage:host coverage ratio was, as seen with the previous system, nearly equal (VIBRANT: 1.02; PHASTER: 1.01) (**Figs. 3E, F**). The effect size of the ratio for the induced samples were also high (VIBRANT: 5.98 and 5.91; PHASTER: 4.92 and 4.88) while the effect size of the control sample ratio was low (VIBRANT: 0.10; PHASTER: 0.05). The results from 5% subsampled reads yielded nearly identical equally determinant values for prophage:host coverage ratios (**Supplemental Figs. 1E, F**) and effect sizes (**Supplemental Figs. 1G, H**).

The third system we tested was *Bacillus licheniformis* DSM13 and its prophages [339]. Here, two prophages were spontaneously activated at 26°C and no control was used for comparison. For VIBRANT, the prophage:host coverage ratios (2.08 and 5.81) as well as the corresponding effect sizes (1.55 and 3.58) were significant (**Figs. 3I, K**). For PHASTER, the prophage:host coverage ratios (1.74 and 1.28) as well as the corresponding effect sizes (0.93 and 0.40) were not significant (**Figs. 3J, L**). The same results for both prediction tools were observed when 5% subsampled reads were used (**Supplemental Figs. 1I-L**).

Although the available control sample size of the three systems and four unique prophages could not designate a true discovery rate with statistical confidence, the controls tested with VIBRANT predictions yielded high accuracy and recall. Specifically, only the *B. krasnovii* prophage in two induced samples yielded a dormant prediction where active was expected.

However, these false-negative results are not entirely unexpected as the default prophage:host coverage ratio for PropagAtE is set very conservatively to 2.0 and can be reduced to 1.75 while maintaining high accuracy. With a ratio cutoff of 1.75, all controls with VIBRANT predictions would have yielded expected results. When PHASTER predictions were used, the false-negative rate for PropagAtE increased considerably indicating that accurate prophage coordinate predictions are essential.

Negative control tests for prophages from isolate genomes

Negative control tests were utilized in order to set threshold boundaries for PropagAtE to identify dormant prophages as well as assess PropagAtE's specificity. Several negative control samples were used for testing in addition to the control samples presented above. Negative controls were considered as those in which a bacterial genome encoding at least one prophage was sequenced in the absence of known prophage induction (i.e., isolate cultures without prophage induction). A total of 19 diverse bacterial genomes encoding 40 predicted prophages by VIBRANT and 37 predicted prophages by PHASTER were used. As before, each system was tested with a set of all reads as well as smaller dataset containing 5% randomly subsampled reads. All sequencing was performed elsewhere (**Supplemental Table S1**). All PropagAtE results for negative control tests can be found in **Supplemental Table S2**.

When using the complete reads sets, all prophages were found to be dormant. Average prophage (1512x to 0.04x) and host (982x to 0.06x) coverages ranged considerably (**Figs. 3M, N**). All prophage:host coverage ratios were below 1.75 (VIBRANT: max 1.55; PHASTER: max 1.54) with the exception of one prophage predicted by PHASTER with a prophage:host coverage ratio of 1.97. However, the effect size of the high prophage:host coverage ratio was only 0.11. All

coverage ratio effect sizes ranged from 2.65 to 0.01 (**Figs. 3O, P**). A total of three prophages predicted by VIBRANT and two prophages predicted by PHASTER had effect sizes greater than 1.75, but the prophage:host coverage ratios were less than 1.55. For the 5% subsampled read results, the prophage:host coverages ranged from 1.55 to 0 and the coverage ratio effect sizes ranged from 2.14 to 0.01. One prophage from each of VIBRANT and PHASTER had an effect size greater than 1.75, but the prophage:host coverage ratio was again less than 1.55 (**Supplemental Figs. 1M-P**).

Given that all prophages were identified as dormant these results suggest that the two metrics, prophage:host coverage ratio and corresponding effect size, function adequately in a check and balance system with each other. Prophages with significantly high prophage:host coverage ratios had insignificant effect sizes, and vice versa. Likewise to the positive control tests, the observed false discovery rate was

zero, though the true accuracy of PropagAtE is likely small but greater than zero. In addition, the negative and positive control tests suggest a prophage:host coverage ratio of 1.75, rather than the conservative default of 2.0, can yield accurate results.

Testing PropagAtE on mock metagenomes

Sequences assembled from complex metagenome samples typically have lower read coverage than those from isolate systems and read mapping is performed in the presence of multiple genomes. We next tested PropagAtE on a mock metagenome consisting of prophages predicted by VIBRANT from 21 unique bacteria from the positive and negative control tests. *Lactococcus lactis* SD96 from the negative controls was not included in favor of *Lactococcus lactis* MG1363 from the positive controls. A total of 21 corresponding read sets, one per host, were

selected and 300k, 100k or 20k paired reads were randomly subsampled per read set and combined to generate the mock metagenome. Thus, three mock metagenomes in total were generated representing 300k, 100k, and 20k subsampled reads per system (**Supplemental Table S3**). The resulting average read coverages of the prophages was 46x, 16x and 3x for the 300k, 100k and 20k subsampled mock metagenomes, respectively. The results from the 300k subsampled reads mock metagenome corresponded to the results from the positive and negative control tests, with 4 active and 36 dormant prophages. A total of 8 prophages with unconfirmed activity status from the positive control hosts were not considered. For the 100k and 20k subsampled reads mock metagenomes, the *B. krasnovii* active prophage was identified as dormant due to insufficient prophage:host coverage ratios (1.75 and 1.70, respectively), and all dormant prophages were accurately identified. This depicts that PropagAtE functions well with combined sequences and partial reads from multiple sources, suggesting the method can work suitably with metagenomes.

Comparing PropagAtE and hafeZ

The software hafeZ [345] similarly utilizes read coverage to identify active prophages. Contrary to PropagAtE, hafeZ does not take in prophage coordinates as input, but rather predicts prophages from a host sequence based on read coverage signatures. Using the hafeZ example *Flavonifractor plautii* host genome and prophages predicted by VIBRANT, PropagAtE correctly identified the expected active prophage with a prophage:host coverage ratio of 3.38 and effect size of 5.98. Conversely, hafeZ was unable to identify any prophages in the positive control datasets presented here. Although, PropagAtE and hafeZ cannot be compared directly due to differing methods of identifying active prophages, these results suggest PropagAtE is better capable of identifying more active prophages than hafeZ.

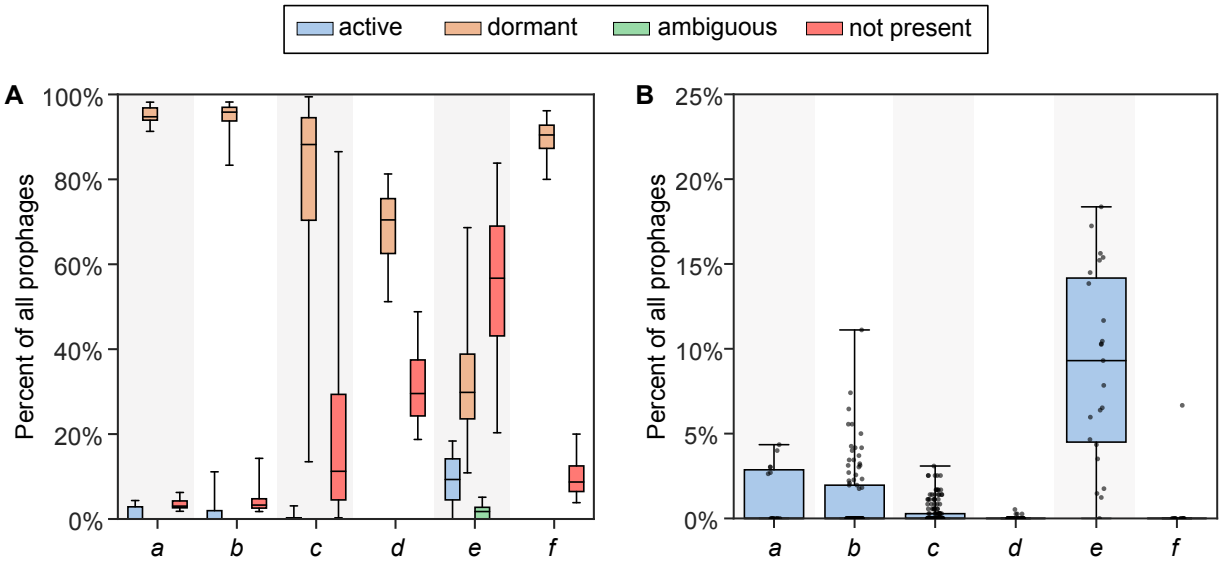


Figure 4. **Percent of prophages by activity category in metagenomic samples**. Five sets of metagenomic samples are compared with (A) all activity categories and (B) only the active prophage category. For (B), each dot represents a single sample. Identifier labels *a-f* on the x-axis correspond to the final column *Label* in **Table 1**.

Applying PropagAtE to identify active prophages in metagenomes

PropagAtE was designed to rapidly assess the activity of prophages in metagenomes in a high-throughput manner. Additionally, PropagAtE can also identify active prophages in genomes of cultivated organisms, irrespective of the manner of prophage induction (i.e., spontaneously or experimentally induced). To validate the broad utility of PropagAtE, we demonstrate its application on 348 metagenomic samples from a variety of environments: adult and infant human gut, murine gut, and peatland soil [85,300,346–349] (**Table 1, Supplemental Table 1**). A total of 349 semi-redundant prophages were identified as active across all samples. Per sample, the percent of prophages that were active ranged from 0% to 18% with a combined average of 1.1% (**Fig. 4**). The murine gut had the most active prophages per sample with an average of 8.9% whereas all human gut samples had a combined average of 1.1%. With a prophage:host coverage ratio of 1.75, the

number of active prophages increased to 402 with a combined average of 1.3%. These results show that for metagenomic samples most prophages identified as integrated into a host genome are dormant or activity is undetectable. All PropagAtE results for metagenomic samples can be found in **Supplemental Table S4**.

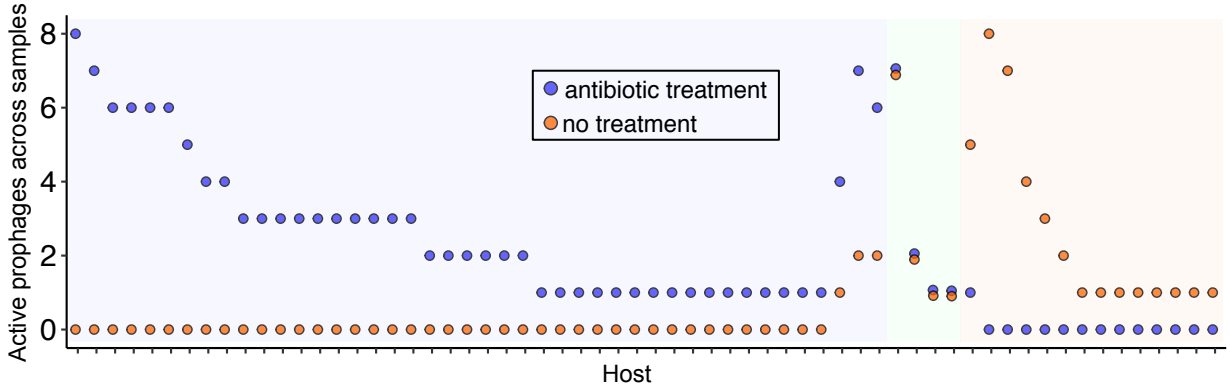


Figure 5. Active prophages identified in infant gut samples. Each host (x-axis) is labeled with two points, one for the total number of prophages identified in antibiotic treatment samples (blue) and one for the total number of prophages identified in no treatment samples (orange). Background highlighting depicts hosts with proportionally more active prophages in antibiotic treatment samples (blue), more active prophages in no treatment samples (orange), or equivalent active prophages in both treatment groups (green).

For metagenome datasets with various conditions (e.g., antibiotic dosage) no significant difference was observed in the total number of active prophages per condition (**Supplemental Fig. S2, Supplemental Table S5**). However, utilizing PropagAtE to identify which sets of prophages are active yielded interesting results. For example, hosts with active prophage populations were compared from the gut of infants given antibiotics compared to infants without antibiotics. A total of 62 host populations with a combined 192 active prophages were compared. Interestingly, a distinct pattern was observed wherein prophage activity was correlated with antibiotic treatment per host population. Generally, a host population had prophage activity in either antibiotic treatment or no treatment, with few host populations having prophage activity uncorrelated with a treatment (**Fig. 5, Supplemental Table S6**). This indicates that although a given prophage or host

population may be found across multiple samples, they may be predominately active in specific treatments.

Estimating prophage activity over time

To further explore the activity of specific prophage populations over time, a sixth set of metagenomic samples was used [304]. This set included human gut fecal samples from three different children with Crohn's Disease. For each individual, four time series samples were taken at approximately days 0, 16, 32 and 54. Among all three individuals, a total of 11 unique prophages were identified across all four time points. None of the 11 prophages were shared between two or more individuals. Therefore, these 11 prophages were found to be consistently present and retained stably over time. All prophage populations encoded hallmark phage proteins, nucleotide replication proteins and lysis proteins, indicating they likely have the ability to activate (i.e., not cryptic). For most populations, genes for integration were also identified (**Supplemental Fig. S3A, Supplemental Table S7**). Furthermore, one prophage population encoded the auxiliary metabolic gene *cysH* for assimilatory sulfate reduction, a metabolic process that can yield hydrogen sulfide, which has been implicated in exacerbating inflammatory bowel diseases such as Crohn's Disease [131,350]. Another prophage population encoded a RhuM family virulence protein. Yet, PropagAtE identified none of these prophages to be active at any time point. This conclusion is important as it suggests that the prophages, in addition to the *cysH* and *rhuM*-like genes, were present but may not have been actively impacting the microbial community at the time of sample collection. Genome alignment of each prophage population yielded 99.8-100% identity with a maximum number of two nucleotide differences between members of a population (**Supplemental Fig. S3B**). The lack of sequence diversification likewise suggests the prophage populations were primarily

dormant over time since active phage genome replication typically results in nucleotide changes. However, the minor nucleotide differences may have resulted from alignment or sequencing error, or from prophage activity between the time points sampled.

Sequencing depth does not correlate with total active prophages

As a final validation test, we examined if the total number of sequencing reads, as an estimation of sequencing depth, had an impact on the total number of active prophages identified. It may be assumed that since PropagAtE relies on read coverage, samples with a greater number of reads would identify disproportionately more active prophages. Using five of the metagenomic sample sets (**Table 1**) we correlated the total number of reads used by PropagAtE to the total number of active prophages identified. Four of the five sets of metagenomic samples yielded near linear, flat trends indicating no correlation between total reads and total active prophages. The fifth set, representing infant gut samples, depicted more of a trend towards a correlation between more reads and more active prophages. However, the trend was not significant (**Supplemental Fig. S4, Supplemental Table S4**).

PropagAtE run time

Efficiency and quick run speed are essential for large-scale metagenomic workflows. PropagAtE was designed to meet the needs of these analyses, such as those with many samples or large file sizes. PropagAtE is likewise scalable for smaller datasets. To show this we estimated the total run time for various isolate and metagenome samples. For isolate samples, run time for PropagAtE analysis was 10-90 seconds with an alignment format file (i.e., BAM format) as the input. For metagenomes, the run time was similar (5-45 seconds) (**Supplemental Table S8**). The

main factor affecting run time is read alignment performed by Bowtie2 which had run times of 1-12 minutes depending on input reads and reference genome sizes. It is important to note that the run time for large read dataset inputs significantly improves when utilizing the multi-threading feature.

Table 1. Summary of metagenomic sample datasets. The environment type, description of the dataset and total number of samples per metagenomic dataset are provided. The final column *Label* corresponds to labeling in **Figure 4**.

Dataset	Description	Samples	Prophages	Hosts	Citation	Label
Human gut (fecal)	Adult individuals with colorectal adenoma, carcinoma, or healthy controls ("CRC")	15	489	484	(57)	*a*
Human gut (fecal)	Adult individuals with Crohn's Disease or healthy controls ("HeQ")	96	2938	2897	(54)	*b*
Human gut (fecal)	Infant individuals given antibiotics or untreated controls ("infant gut")	139	356	333	(55)	*c*
Peatland (soil)	Peatland soil cores of bog, fen and palsa environments ("soil")	75	379	375	(22, 56)	*d*
Murine gut (fecal)	Virome fraction samples from the murine gut ("murine gut")	23	1308	1292	(53)	*e*
Human gut (fecal)	Time series of adult individuals with Crohn's Disease ("IjazUZ")	12	155	153	(58)	*f*

Discussion

Phages are key contributors to microbiome dynamics in essentially all environments on Earth [5,9,12,35,37,84,122,351]. With the availability of high-throughput sequencing and newly developed software tools we have the ability to identify and study these diverse phages [115,117,295,296]. This includes both strictly lytic phages as well as integrated prophages. However, little emphasis has been placed on identifying which populations of identified prophages are actively replicating as opposed to existing in a dormant or cryptic stage of infection.

Here we have presented the software tool PropagAtE for the estimation of activity of integrated prophages using statistical analyses of read coverage. Although the concept of using

read coverage to predict prophage activity is not new [339], PropagAtE is the first benchmarked implementation of the method into an automated software for use with large datasets, such as metagenomes. PropagAtE functions by quantifying the relative genome copy ratio between a prophage region compared to a corresponding host region. Only prophages that have activated and begun propagation (e.g., genome replication and virion assembly) will yield prophage:host ratios sufficiently greater than 1:1. The prophage:host genome copy ratio, estimated by using read coverage ratios, as well as the ratio's effect size are used to classify a prophage as active or dormant. We provide evidence to show that PropagAtE is fast, sensitive, and accurate in predicting prophages as active versus dormant and have applied the method to various metagenome samples.

Identifying which prophage sequences are active versus dormant in a sample provides several benefits. Namely, assuming that all identified prophages are active is an overestimation and will lead to a misrepresentation of the *in situ* dynamics of a microbial community. For example, we show here that 11 unique prophages identified in human gut samples from the same individual over time may not necessarily be active when identified. The most accurate representation of the prophages is to conclude that their effect on the resident microbial communities likely occurred at a time point not sampled or that the prophages were consistently dormant. Another benefit includes making accurate conclusions on the role of host bacteria in a given sample. Foremost, prophages can be responsible for the virulence of multiple human pathogens, such as *Clostridioides difficile*, *Clostridium botulinum*, *Staphylococcus aureus,* and *Corynebacterium diphtheria* [34,352–356]. Although some virulence effects are present during prophage dormancy and expression of specific genes, many require activation of the prophage. In addition to virulence, bacteria actively infected by a phage can have a modified metabolic landscape compared to bacteria uninfected or harboring a dormant prophage. Several examples

include the phage-directed regulation of sulfur, carbon, nitrogen and phosphorus metabolism in various cyanobacteria and enterobacteria [66,69,70,75]. This distinction is vital when assessing the role of the microbial community in an environment. Related to this, activity can provide context to any auxiliary metabolic genes identified on the prophage genome, such as *cysH* for assimilatory sulfate reduction described here. In the human gut specifically, identifying phage-encoded genes for sulfur metabolism may have important implications for the health of the gastrointestinal tract and a phage's role in the manifestation or perturbation of diseases [13,351]. If a prophage encoding an auxiliary metabolic gene is identified, determining the stage of infection of the prophage can provide context to the effect of the auxiliary metabolic gene.

It is important to point out several unavoidable caveats to the implementation of PropagAtE. First, accurate prophage:host genome copy ratio estimations are inhibited if the sample is size fractionated before sequencing. For example, many aquatic samples are size fractionated by filtering onto a 0.2-micron filter. In these cases, only pre-lysis infections will be picked up by read coverage because the genomic content present in released virions will likely pass through the 0.2-micron filter. Second, not all prophages exist as integrated sequences, such as those that are episomal. Prophages that are episomal do not have attached host sequence and therefore cannot have prophage:host read coverage compared in a one-to-one manner, and for metagenomes cannot have accurate host prediction. This also applies to prophages that do not assemble as integrated components of a host scaffold. However, it is worth noting that for integrated prophages PropagAtE functions whether the host region flanks the prophage on one or both sides. Third, though not verified, is that inactive prophages may be more likely to assemble with a host scaffold. Since active prophages lyse their host and potentially degrade their host's genome, more activity of a prophage may lead to a lower probability of assembling as an integrated prophage. Fourth,

induction of prophages within a host population may occur asynchronously and lead to consistent activity with low prophage:host coverage ratios, causing activity to be missed. Fifth, some host populations may include some members that encode a prophage and some members that do not. In the latter example, the prophage:host ratio is initially skewed to less than one, making it more likely for PropagAtE to miss activity. Due to the caveats presented, PropagAtE is intended to be used for identifying active prophage sequences rather than assessing the total number or fraction of prophages that are active in a sample. In this context PropagAtE performs with little to no observed error. Finally, PropagAtE has been developed and tested using short read sequencing data and is not yet suitable for long read analyses.

Overall, our results demonstrate that PropagAtE will facilitate the accurate characterization and study of viruses in microbiomes and nature. Examples of future applications of PropagAtE include the exploration of prophages in human health and disease, detection of environmental and chemical triggers for induction of prophages, phage therapy research (for disqualifying prophages), and in environmental systems research.

Methods

Datasets used for control tests

All datasets, genomes and reads, used for positive and negative control tests were acquired from publicly available datasets on NCBI databases [207,210]. See **Supplemental Table S1** for details of studies and accession numbers. VIBRANT (v1.2.1) [117] and PHASTER (accessed December 2021) were used for identification and annotation of all prophages. Only VIBRANT was used for identification of prophages from metagenomes. For the mock metagenome, reads were randomly subsampled using seqtk (v1.3-r106, sample) (https://github.com/lh3/seqtk).

Dependencies and equations

Bowtie2 (v2.3.4.1) [225] was used for read alignment. Samtools (v1.11) [340] and PySam (https://github.com/pysam-developers/pysam) were used for manipulation, conversion, and reading of SAM and BAM alignment files. To calculate coverage, aligned reads are filtered according to the percent identity alignment, as calculated by subtracting number of gaps g and the number of mismatches m in the alignment from the length of the alignment l, and then dividing by l.

$$\text{percent identity alignment} = \frac{l - g - m}{l} \cdot 100\%$$

Cohen's *d* metric is used to calculate the effect size of prophage:host coverage ratios. Cohen's *d* [357] is calculated using the following equation where $\bar{X}_{host}$ and $\bar{X}_{prophage}$ are the average read coverages of the host and prophage regions, and S_{host} and $S_{prophage}$ are the standard deviations of the coverages:

$$d = \frac{\bar{X}_{host} - \bar{X}_{prophage}}{\sqrt{\frac{S_{host}^2 + S_{prophage}^2}{2}}}$$

Metagenome assembly and analyses

Metagenomes for the murine gut microbial fraction samples were assembled in this study. Details of raw read sets from murine gut samples used for assembly can be found in **Supplemental**

Table S1. SPAdes (v3.12.0) [90] was used for genome assembly (--meta -k 21,33,55) and the resulting best scaffold assemblies were retained. The human infant gut and peatland soil metagenomes were assembled previously in their respective studies [85,347,348]. Both human adult gut metagenomes were assembled by Pasolli et al. [160].

For the human gut time series samples integrated prophages were predicted using VIBRANT (v1.2.1). To check for integrated prophage sequences that were not assembled with a host scaffold, integrated prophages were compared to all identified phages using dRep (v2.6.2, dereplicate –ignoreGenomeQuality -sa 90 -pa 90) [219]. Identical, non-integrated phage sequences were considered as a part of the same prophage population. Genome alignments were performed using progressive Mauve (v1.11, default settings) [358].

Visualization

Geneious Prime 2020.1.2 was used for visualization of example read coverage values. R package 'ggplot2' and Python packages Matplotlib and Seaborn were used for visualization of graphs [228,270].

Setting default thresholds for PropagAtE

PropagAtE has several, variable settings and thresholds that can be set by the user: percent identity of aligned reads, masking of coverage values at genome/scaffold ends, minimum prophage:host coverage ratio, minimum Cohen's *d* effect size, minimum average coverage of the prophage, and minimum breadth of coverage of the prophage. In addition, PropagAtE requires that all prophage and host sequences must each be at least 1 kb in length.

Percent identity read alignment is used for more accurate read alignment processing. This setting is meant to be sensitive for accurate read alignment while allowing for minor errors (default: 97%). Another coverage metric is masking of coverage values at genome/scaffold ends. This setting is particularly important for metagenomic scaffolds that likely represent partial sequences. For this metric, a generalized length of 150 base pairs is used to mask (i.e., not consider for calculation) the respective number of coverage values from each scaffold end in order to account for lower coverage values at partial scaffold ends.

The final four settings are used for determination of prophage activity and significance: The most important threshold is the prophage:host coverage ratio, which is set to 2.0 by default and can be reduced to 1.75 for increased sensitivity. The default was selected to be as close to the minimum requirement for designating true active prophages as active in control tests while maintaining a significant gap from true dormant prophages in order to reduce false positive identifications. Finally, Cohen's *d* effect size setting is set to 0.70 which falls in the general range of "medium" significance [357]. This threshold is useful for contextualizing prophage:host coverage ratios, especially for high-coverage genomes/scaffolds. Again, the default was selected according to control tests for reducing false positive identifications. The thresholds for minimum coverage (default: 1.0) and minimum breadth (default: 0.50) of prophage regions are used to ensure that only prophages that are likely to be present in the sample (i.e., sufficient coverage) are considered in analyses.

Data Availability

The PropagAtE software and associated files are freely available as a Python package at https://github.com/AnantharamanLab/PropagAtE. All isolate and metagenome genomic

sequences and reads used in this study are publicly available; see **Supplemental Table S1** for details. Additional details of relevant data are available on request.

Acknowledgements

We thank the University of Wisconsin—Office of the Vice Chancellor for Research and Graduate Education, University of Wisconsin—Department of Bacteriology, and University of Wisconsin—College of Agriculture and Life Sciences for their support. We also thank Z. Zhou, A. Adams, and R. Salamzade for their helpful feedback and discussions. K.K. was supported by a Wisconsin Distinguished Graduate Fellowship Award and a William H. Peterson Fellowship Award from the University of Wisconsin-Madison. This research was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM143024.

Author Contributions

K.K and K.A designed the study, performed all analyses and interpretation of data, and wrote the manuscript. All authors have reviewed and approved the final manuscript.

Chapter 6: vRhyme enables binning of viral genomes from metagenomes

Kristopher Kieft[1,2], Alyssa Adams[1,3], Rauf Salamzade[2,4], Lindsay Kalan[4,5], and Karthik Anantharaman[1]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA
[2]Microbiology Doctoral Training Program, University of Wisconsin–Madison, Madison, WI, USA
[3]Computation and Informatics in Biology and Medicine, University of Wisconsin–Madison, Madison, WI, USA
[4]Department of Medical Microbiology and Immunology, University of Wisconsin–Madison, Madison, WI, USA
[5]Department of Medicine, University of Wisconsin–Madison, Madison, WI, USA

Publication:
Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. *vRhyme enables binning of viral genomes from metagenomes*. *bioRxiv*, 2021.12.16.473018 doi:10.1101/2021.12.16.473018.

All supplementary figures, tables, and files are available at the following Figshare repository:
https://figshare.com/projects/Kristopher_Kieft_PhD_Dissertation/136427

Abstract

Genome binning has been essential for characterization of bacteria, archaea, and even eukaryotes from metagenomes. Yet, few approaches exist for viruses. We developed vRhyme, a fast and precise software for construction of viral metagenome-assembled genomes (vMAGs). vRhyme utilizes single- or multi-sample coverage effect size comparisons between scaffolds and employs supervised machine learning to identify nucleotide feature similarities, which are compiled into iterations of weighted networks and refined bins. To refine bins, vRhyme utilizes unique features of viral genomes, namely a protein redundancy scoring mechanism based on the observation that viruses seldom encode redundant genes. Using simulated viromes, we displayed superior performance of vRhyme compared to available binning tools in constructing more complete and uncontaminated vMAGs. When applied to 10,601 viral scaffolds from human skin, vRhyme advanced our understanding of resident viruses, highlighted by identification of a Herelleviridae vMAG comprised of 22 scaffolds, and another vMAG encoding a nitrate reductase metabolic gene, representing near-complete genomes post-binning. vRhyme will enable a convention of binning uncultivated viral genomes and has the potential to transform metagenome-based viral ecology.

Introduction

Viruses and bacteriophages (collectively termed viruses) are pervasive members of essentially all ecosystems. Viruses form a continuum of symbiotic interactions with their hosts, from lethal parasitism to essential mutualism[281,359,360]. These interactions are known to impact biogeochemical and nutrient cycling processes, human health, infrastructure and industries, and ecosystem community dynamics[21,36,70,351]. As a result of the rising interest in viromics, the

previously unknown members of the virosphere, the range in the encoded genetic potential of viruses, known viral diversity, and limits of viral genome sizes have been continuously expanding[9,60,71,180,361].

Metagenomic sequencing can be a mechanism to identify, recognize, understand, and even harness the information encoded on viral genomes. Most metagenomes will assemble into many short fragments (scaffolds or contigs) representing partial genome sequences. The process of binning is employed to group scaffolds into a putative genome, termed a metagenome-assembled genome (MAG). With the information encoded by a MAG, rather than individual scaffolds, stronger inferences of metabolic potential, phylogenies, taxonomy, and community interactions can be generated[309].

Conversely, viral scaffolds are typically not binned. Handling complex and often enigmatic viral scaffolds in metagenomes often poses computational challenges unique from microbes. One justification to not bin viruses is that their genomes are small relative to cellular organisms and the assumption that most scaffolds represent the majority, or the entirety, of an identifiable genome. For dsDNA viruses, the target of most viral metagenomes, genome sizes will have a general range of 20 kb – 200 kb, with the largest of viruses being 500 kb – 2000 kb. Since the majority of scaffolds in most assembled metagenomes are below 20 kb in length, it can be estimated that a single scaffold likely will not represent an entire viral genome. In fact, benchmarks have shown that viruses often do not assemble into a single scaffold[144,362]. Another difficulty with binning viral genomes is that viruses do not encode universal single copy or marker genes, making a standardized approach for all viruses difficult to create.

Many software tools have been developed for binning bacterial, archaeal, and eukaryotic metagenomic scaffolds into MAGs[92,93,95,96,363–368]. These tools employ a wide range of

methodologies, mainly focusing on tetranucleotide frequencies and read coverage abundance variance comparisons between scaffolds. A significant portion of the tools tailored to bacteria and archaea also rely on identifying microbial single copy genes to inform the construction of bins along with completeness and contamination estimates. Some tools for binning microbes are suitable for binning viruses due to their independence from microbial single copy gene analysis, namely MetaBat2, VAMB, CONCOCT, and BinSanity. MetaBat2 uses a composite scoring system based on the geometric mean of tetranucleotide frequencies and coverage abundance of individual scaffolds to generate bins according to a weighted graph clustering algorithm[92]. VAMB implements unsupervised deep learning variational autoencoders based on individual scaffold tetranucleotide frequencies and coverage abundance to generate bins by iterative medoid clustering[95,119]. CONCOCT uses tetranucleotide frequencies and coverage abundance, reduced by multidimensional reduction, to cluster scaffolds into bins with Gaussian mixture models[94]. BinSanity uses affinity propagation clustering based on coverage abundances to bin scaffolds, followed by bin refinement using tetranucleotide frequencies and GC content[367]. Despite the abundance of tools for binning bacteria and archaea, there is a conspicuous dearth of tools available for binning viruses. Only one tool, CoCoNet[118], has thus far been developed for binning viral genomes from metagenomes (viral MAGs, or vMAGs). CoCoNet implements an unsupervised deep learning neural network to identify shared tetranucleotide and coverage abundance patterns between scaffold pairs, followed by graph clustering of potential pairs into bins[118].

Here, we present vRhyme, a software tool that incorporates supervised machine learning based classification of diverse sequence feature compositions as well as read coverage abundance effect size comparisons to generate weighted networks of bins. vRhyme leverages unique features of viral genomes to optimize and refine the binning of vMAGs, including overcoming the lack of

single copy genes by scoring protein redundancy based on the observation that viruses seldom encode redundant genes. vRhyme is capable of binning viruses from diverse families, host and source environment affiliations, varying states of genome fragmentation, and wide ranges of genome lengths. In benchmarking vRhyme, we show that it is fast, inclusive, and accurate in binning viral scaffolds, with low computational demands, in synthetic and natural metagenomes compared to other binning software. When applied to human skin metagenomes, we show that vRhyme enabled a more comprehensive analysis of shared viruses and viral features across a cohort of individuals, and likely better recapitulated natural systems. vRhyme is implemented in Python and is freely available for download at https://github.com/AnantharamanLab/vRhyme.

Methods

Coverage processing

The input for read coverage information is variable: paired or unpaired short reads, SAM alignment file, BAM alignment file, or a pre-calculated coverage table. For short reads input, reads will be aligned to input scaffolds using either Bowtie2[104] or BWA[103]; Bowtie2 is run with the parameters --no-unal --no-discordant, the latter being for paired reads only, and BWA is run with the mem algorithm. All reads should be quality filtered before being used as input. The resulting SAM alignment file, or an input SAM alignment file, will be converted into BAM format using Samtools[340]. BAM alignment files, either generated by the vRhyme pipeline or as user input, will then be processed. As such, any input combinations of short reads, SAM or BAM alignment files are compatible. BAM alignment files, if not already provided as input, are sorted and indexed using Samtools.

The Python package Pysam (https://github.com/pysam-developers/pysam) is then used to fetch aligned records within sorted and indexed BAM alignment files for processing and coverage calculations. First, aligned reads are filtered according to the percent identity alignment, as calculated by the sum of the number of gaps g and the number of mismatches m in the alignment divided by the length of the alignment l. The default is a 97% identity alignment.

$$\text{percent identity alignment} = \frac{l - g - m}{l} \cdot 100$$

Aligned reads passing the set threshold are used to calculate the total coverage of each nucleotide base per scaffold, inclusive of bases with a coverage of zero. Finally, the coverage values at the terminal ends of scaffolds are masked to increase coverage fidelity by considering erroneous read alignment at partial scaffold ends. The default is to ignore all coverage values within the first and last 150 bp of the scaffold. The average and standard deviation of coverage per scaffold is calculated according to respective, individual base coverages. All alignment filtering and coverage calculations are handled natively within vRhyme. This final step yields a coverage table comprised of the average and standard deviation of coverage per scaffold per input sample. This coverage table, or a user-generated table of the same format, can be used as input for vRhyme in place of reads or SAM/BAM alignment files.

Next, scaffold coverages across all *k* samples are pairwise compared using the effect size of coverage differences. First, all average coverages are increased by a pseudo-count of 0.1 to avoid coverages of zero (pseudo-counts are excluded from coverage table). Effect size is calculated by the Cohen's *d* effect size metric equation[357]. Cohen's *d* is calculated as follows, where $\bar{X}_i$ and

$\bar{X}_j$ are average read coverages and σ_i and σ_j are standard deviations of the coverages for a scaffold pair *i* and *j*:

$$d_{k,i,j} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{\sigma_i^2 + \sigma_j^2}{2}}}$$

For each pairwise comparison, an effect size value d_k is generated per sample *k* across all samples *n*. Values exceeding the effect size threshold, set by vRhyme presets, generate an additive penalty weight *p*. The average effect size across all samples $\bar{X}_d$, with any added penalties, is normalized to the number of input samples, yielding a normalized effect size d', which considers higher statistical power to more sample comparisons:

$$\bar{X}_{d,i,j} = \frac{\sum_{k=1}^{n} d_{k,i,j}}{n} + p_{i,j}$$

$$d'_{i,j} = \frac{\bar{X}_{d,i,j}}{\log_{10}(n) + 1}$$

The normalized and penalized d' values are compared to a normalized preset effect size threshold and all pairwise comparisons passing the set criteria are considered as co-occurring by coverage. Any scaffold not found to co-occur with another is discarded. For computational efficiency, a pre-filter is applied where only the best (i.e., lowest d') n pairs per individual scaffold are retained, where n is '--max_edges' multiplied by 3.

Nucleotide processing

All co-occurring scaffolds by read coverage are compared by seven nucleotide content metrics. The pairwise distance calculations per metric are used as inputs to supervised machine learning models for classification. All nucleotide features and distances are calculated natively within vRhyme.

The first feature, codon usage (CU), is calculated from nucleotide open reading frames (i.e., genes). Predicted genes can be used as input, otherwise vRhyme will automate prediction using Prodigal[98] (-m -p meta). In-frame trinucleotide counts c for each of the 64 codons k (step of 3 bases) along a scaffold are divided by the total count of observed codons. The final codon, if representing a stop, is ignored. Counts are inclusive of zero counts but exclusive of ambiguous (e.g., N) bases. The following yields a CU frequency vector F_i for each codon k in scaffold i.

$$F_k = \frac{c_k}{\sum_{k=1}^{64} c_k}$$

$$F_i = (F_1, F_2, F_3 \ldots F_k)$$

The next three features (GC content, CpG content, and GC-skew) are calculated per scaffold from individual scaffold bases. GC content N_{gc} is calculated by the sum of all G and C bases, divided by the sum of all bases (A, T, C and G). CpG content N_{cpg} is calculated by the sum of all CG di-nucleotides per scaffold (step of 1 base) divided by the sum of all bases. GC-skew N_{skew} is calculated by subtracting the total of C bases from the total G bases, divided by the sum of G and C bases.

$$N_{gc} = \frac{C + G}{C + G + A + T}$$

$$N_{cpg} = \frac{CG}{C + G + A + T}$$

$$N_{skew} = \frac{G - C}{G + C}$$

The last three features – relative tetranucleotide frequency (RTF), tetranucleotide usage deviation (TUD) and tetranucleotide zero'th order Markov method (ZOM) – are calculated from whole scaffold tetranucleotide frequencies (step of 1 base) of the forward and reverse strands[369]. A total of 136 possible tetranucleotides are considered after combining identical, reverse complement and palindromic sequences. Counts are inclusive of zero counts but exclusive of ambiguous (i.e., N) bases.

For RTF, all counts t for each of the 136 tetranucleotides k along a scaffold are divided by the total count of observed tetranucleotides. The following yields a tetranucleotide frequency vector T_i for each tetranucleotide k in scaffold i.

$$T_k = \frac{t_k}{\sum_{k=1}^{136} t_k}$$

$$T_i = (T_1, T_2, T_3 \ldots T_k)$$

For TUD, expected nucleotide frequencies E are first calculated by dividing the count of each base b by the sum of all bases in the scaffold. Next, observed counts per base O_b per tetranucleotide k are calculated by the sum of each base inclusive of zero counts. For each unique tetranucleotide, expected frequencies per base are raised to the power of observed frequencies multiplied by two to yield a deviation value D_b per base. The deviation values for all four bases are multiplied the count of total observed tetranucleotides and the count of the given tetranucleotide to yield a TUD value per tetranucleotide. The following yields a TUD frequency vector TUD_i for each tetranucleotide k in scaffold i.

$$E_b = \frac{b}{C + G + A + T} \text{ for } b = \text{A, T, C, G}$$

$$O_b = \sum_{k=1}^{4} b \quad \text{for } b = \text{A, T, C, G}$$

$$D_b = E_b^{(2 \cdot O_b)} \quad \text{for } b = \text{A, T, C, G}$$

$$TUD_k = D_A \cdot D_T \cdot D_C \cdot D_G \cdot \sum_{k=1}^{136} t_k \cdot t_k$$

$$TUD_i = (TUD_1, TUD_2, TUD_3 \ldots TUD_k)$$

For ZOM, the same expected E_b nucleotide frequencies per base b are used. For each tetranucleotide k, the count t of the given tetranucleotide is divided by the product of each of the

present tetranucleotide's bases' expected frequencies to yield a ZOM frequency vector ZOM_i for each tetranucleotide k in scaffold i.

$$ZOM_k = \frac{t_k}{E_{b_1} \cdot E_{b_2} \cdot E_{b_3} \cdot E_{b_4}}$$

$$ZOM_i = (ZOM_1, ZOM_2, ZOM_3 \ldots ZOM_k)$$

Pairwise distance calculations for GC, CpG and GC-skew are made by the absolute value difference in the respective metric's content between two scaffolds. For example, the following is the pairwise distance P_{GC} in GC content between scaffolds i and j.

$$P_{i,j} = |GC_i - GC_j|$$

Pairwise distance calculations for CU, RTF, TUD and ZOM are made by cosine distances. For each value v_i and v_j, corresponding to the same tetranucleotide k, in frequency vectors of scaffolds i and j, with vector averages of $\bar{V}_i$ and $\bar{V}_j$, cosine similarity $S_{i,j}$ is calculated. Cosine distances between two scaffolds are calculated for CU, RTF, TUD and ZOM individually.

$$S_{i,j} = \frac{\sum_{k=1}^{n}(v_{i_k} \cdot v_{j_k})}{\sqrt{\left(\sum_{k=1}^{n}(v_{i_k} \cdot \bar{V}_i)\right)^2 \cdot \left(\sum_{k=1}^{n}(v_{j_k} \cdot \bar{V}_j)\right)^2}}$$

The result of distance calculations is a vector $M_{i,j}$ of length seven for each pairwise comparison between scaffolds i and j.

$$M_{i,j} = (N_{GC}, N_{CpG}, N_{skew}, S_{CU}, S_{RTF}, S_{TUD}, S_{ZOM})$$

Machine learning model training and testing

NCBI databases (RefSeq[207] and Genbank[210], release July 2019) were queried for "prokaryotic virus" and genomes greater than 10 kb in length were retained. In addition, the IMG/VR database (release July 2018)[141] was downloaded, and sequences were limited to a minimum length of 10 kb. For the IMG/VR dataset, VIBRANT[117] (v1.2.1, -virome) and CheckV[370] (v0.6.0) were used to obtain circular and/or complete sequences. The resulting NCBI and IMG/VR datasets were dereplicated by 95% identity using the method described here (--derep_only --derep_id 0.95 --frac 0.70 --method longest) and combined, resulting in a total of 11,881 putatively complete genomes. The sequences representing complete genomes in the combined dataset were split into non-overlapping fragments of 15 kb with a minimum length of 10 kb. A total of 39,105 fragments were generated for training and testing machine learning models, with 38,732 represented in the training and 30,618 represented in the testing datasets (**Supplementary Figure 1a**).

The machine learning models were generated based on the $M_{i,j}$ vectors described above using the generated 39,105 genome fragments. Filtering of pairwise comparisons before training and testing was made according to vRhyme default parameters (--max_gc 0.20 --min_kmer 0.60). The pairwise comparison matrix was split 75:25 for training and testing, respectively. Fragment pairs were labeled as "same" or "different" for supervised machine learning according to if the

paired fragments originated from the same or different source genomes. An equal number (69,632) of "same" and "different" pairs were used for training by randomly dropping excess "different" comparisons. For testing, a set of 38,685 "different" and 7,736 "same" pairs were used. There were no redundant pairs between the training and testing datasets.

Scikit-Learn (v0.24.2)[299] was used to generate machine learning models using a grid search approach to optimize parameters. Several models and algorithms were considered, including MLPClassifier, ExtraTrees, KNeighbors, SVC, Gradient Boost, Decision Tree and Random Forest classifiers. Iterative training and testing yielded MLPClassifier (alpha=0.001, beta_1=0.7, beta_2=0.8, hidden_layer_sizes=(5,25,50,75,100,100,75,50,25,5), learning_rate_init=0.0001, max_iter=1250, n_iter_no_change=15, tol=1e-08) and ExtraTreesClassifier (max_depth=10, max_features=7, n_estimators=1500) as the most robust.

Machine learning and network processing

Each scaffold pair is classified by the two machine learning models separately to yield two probability values of "same", one per model. The probability values are averaged to yield $\bar{p}$. Any pair with $\bar{p}$ below the preset threshold is discarded. Then, d' calculated previously for the pair is divided by $\bar{p}$ to yield a network edge weight w.

$$w = \frac{d'}{\bar{p}}$$

Any pair with w below the preset threshold is retained for network clustering. As before, for computational efficiency, only the best (i.e., lowest w) n pairs per individual scaffold are retained, where n is '--max_edges'. Weighted networks, representing unrefined bins, are created where each node is a scaffold and each edge is a weighted connection between paired scaffolds.

Networks are refined using MiniBatchKMeans implemented in Scikit-Learn with the following parameters: n_clusters=s+1, batch_size=h, max_iter=100, max_no_improvement=5, n_init=5. Batch size h is 25% of the number of nodes with a minimum of 2 and maximum of 100. The number of clusters s is defined by the number of nodes with a clustering coefficient value below the preset constant 0.36 but not 0. For each node i, the clustering coefficient U_i is calculated as follows, where L_i is the degree of the node and R_i is the number of edges between the neighbors of i:

$$U_i = \frac{2 \cdot L_i}{R_i \cdot (R_i - 1)}$$

Refined networks are split into distinct, separate networks according to s. Here, each connected network represents a putative bin.

Score processing

Each binning iteration is given a score I according to protein redundancy, total bins, and the number of scaffolds binned. To calculate protein redundancy, all proteins within a bin are clustered using Mmseqs2[271] (linclust --min-seq-id 0.5 -c 0.8 -e 0.01 --min-aln-len 50 --cluster-mode 0 --seq-id-mode 0 --alignment-mode 3 --cov-mode 5 --kmer-per-seq 75). Any proteins clustered within a bin, excluding those along the same scaffold, are considered redundant. The iteration with the maximum score is selected as the final representative. I is calculated as follows:

$$I_r = \sum_{bin=1}^{n} \frac{proteins\ clustered_{bin} - number\ of\ clusters_{bin}}{total\ proteins_{bin}}$$

$$I_s = \frac{scaffolds\ binned}{input\ scaffolds}$$

$$I_b = \frac{number\ of\ bins}{scaffolds\ binned}$$

$$I = I_s - {I_b}^2 - (3 \cdot \sqrt{2 \cdot I_r})$$

Dereplication

vRhyme implements Nucmer[264] and MASH[263] for the dereplication of scaffolds. First, scaffolds are roughly grouped using MASH (sketch -k 31 -s 1000; dist) to reduce the pairwise comparison space. Next, all possible pairs of scaffolds within each resulting group are aligned using Nucmer (-c 1000 -b 1000 -g 1000). Regardless of the comparison method ('--method'), any pair of scaffolds with 100% identity over 100% coverage are first reduced to the longest representative. For all percent coverage calculations in dereplication, coverage is of the shortest scaffold. For '--method longest' the longest scaffold in pairs meeting the set percent identity (e.g., 97%) and percent coverage (e.g., 60%) thresholds is taken as the representative. For '--method composite', scaffold pairs meeting the percent identity and percent coverage thresholds are joined over the region of sequence overlap to yield artificially chimeric scaffolds. Any alignments exceeding the sensitivity values for merging over complex alignments, such as low identity scaffold ends without overlap, are not joined. After scaffold pairs are joined, identical cycles of MASH, Nucmer and composite joining are completed until no further alignments are detected. For all methods, reverse complement sequence alignments are considered and adjusted accordingly.

Performance validation datasets and metrics

Scaffolds used to benchmark performance were acquired from nine separate publicly available datasets derived from eight unique metagenomes (one metagenome was split into two separate datasets). The metagenomes were acquired from marine[82,371], freshwater[372–374], human gut[300], and soil environments[348,375]. Details on the studies, scaffolds, reads, and accession numbers can be found in **Supplementary Table 1** Each dataset was processed separately. First, VIBRANT (v1.2.1) was used to predict viruses. From these viruses, VIBRANT and CheckV were used to identify circular scaffolds representing complete genomes. Next, scaffolds were dereplicated by 97% identity using the method described here (--derep_only --derep_id 0.97 --frac 0.70 --method longest). The non-redundant scaffolds were randomly fragmented into sequences ranging from 2 kb to 20 kb in length. A total of 999 scaffolds (i.e., putatively complete genomes) were used to generate 4,324 fragments of at least 2 kb in length. Full benchmarking was performed on the 4,324 fragments and validation of complete genome binning was performed on the 999 scaffolds representing complete genomes (**Supplementary Figure 1b**). Only 255 of the performance benchmarking fragments had significant sequence similarity to fragments used to train the machine learning models (**Supplementary Figure 1c**).

Since the circular scaffolds (sources) were estimated to be complete genomes, any of the fragments originating from the same source were expected to create a single bin, bins containing fragments from multiple sources were considered as contaminated, fragments from the same source in different bins were considered as split genomes, and fragments representing an entire source (singletons) were not expected to bin. The following equations are for genome- (source) and bin-based performance metrics, where B_e is the expected number of bins (i.e., sources with at

least two fragments), B_g is the number of bins generated, G_e is the expected number of binned fragments (i.e., fragments representing B_e sources), B_o is the total number of bins containing a single source, G_t is the total number of fragments binned, G_b is the number of unique sources binned, G_o is the number of sources contained in a single bin, G_s is the total number of singletons, and G_p the number of binned singletons.

$$binned\ singletons = \frac{G_p}{G_s}$$

$$genome\ recall = \frac{G_t - G_s}{G_e}$$

$$genome\ precision = \frac{G_o - G_p}{G_b}$$

$$genome\ splitting = \frac{G_b - G_o}{G_b}$$

$$bin\ precision = \frac{G_o}{B_g}$$

$$bin\ contamination = \frac{B_g - B_o}{B_g}$$

$$genomes\ total = \frac{G_b}{B_e}$$

$$bins\ total = \frac{B_g}{B_e}$$

$$bins{:}\ genomes = \frac{B_g}{G_b - G_s}$$

$$genomes{:}\ bins = \frac{G_b - G_s}{B_g}$$

$$genomes\ score = \frac{2 \cdot (G_o - G_p)}{(2 \cdot (G_o - G_p)) + G_p + G_b - G_o}$$

$$bins\ score = \frac{2 \cdot G_o}{(2 \cdot G_o) + (B_g - B_o)}$$

To validate binning further, each pairwise connection between fragments within a bin was evaluated according to each fragment's nucleotide length. These standard performance metrics were evaluated per bin using true positive TP, true negative TN, false positive FP, and false negative FN connections. The following equations are for pairwise nucleotide-based performance metrics:

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

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

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$specificity = \frac{TN}{TN + FP}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

Performance benchmarking

The performance of vRhyme (v1.0.0) was compared to MetaBat2[92] (v2.12.1, -s 4000 -m 2000), CONCOCT[94] (v1.0.0, -l 2000), VAMB[95] (v3.0.2, -i 2 -m 2000 -t 40), CoCoNet[118] (v1.0.0, --min-ctg-len 1000 --min-prevalence 1), and BinSanity[367] (v0.5.4, -x 2000). Additional binning tools, namely MaxBin2[93], MyCC[363], SolidBin[364] and DASTool[366], perform microbial single copy gene analysis and were not applicable, or did not function, for viruses. For VAMB, the starting batch size had to be adjusted to accommodate the relatively small size of the input datasets, and all but three datasets failed to run. The coverage tables for each of the tools were generated from sorted BAM files using each tool's respective method, except for VAMB for which the same coverage table as MetaBat2 was used. The sorted BAM files were generated using Samtools (v1.13) with reads quality filtered by Sickle[275] (v1.33) aligned by Bowtie2 (v2.3.5.1, --no-unal --no-discordant).

Metagenomic datasets and analyses

Publicly available metagenomes from marine[10], agricultural soil[106], and human skin[376] environments were used. Details on the studies, reads used, and accession numbers can be found in **Supplementary Table 1**. Viruses were predicted from each metagenome using VIBRANT and only the identified virus scaffolds were binned using vRhyme. For the human skin datasets, 270 metagenomes from a cohort of 34 individuals with eight body sites per individual were used (antecubital fossa (Af), alar crease (Al), back (Ba), nare (Na), occiput (Oc), toe-web space (Tw), umbilicus (Um), and volar forearm (Vf)). Reads were filtered for quality, adapters, and host-contamination as described previously[376] using fastp[377] (v0.21.0, --detect_adapter_for_pe) and KneadData (v0.8.0). MegaHit[378] (v1.2.9) was used to generate individual metagenomic assemblies for each sample, corresponding to the microbiome of a particular body site for a specific participant at a given timepoint. After predicting viruses, all viruses per body site were combined and dereplicated (--method longest) before binning.

It is important to note that for bins, scaffolds had to be linked with Ns in order to run CheckV analysis since there is no mode to input bins. For all benchmarking using CheckV, the tool was modified to run Prodigal with the -m flag to accommodate linking vMAGs and not predicting open reading frames across the appended strings of Ns connecting scaffolds. For taxonomy of the validation dataset, a publicly available custom reference database of NCBI viruses was used as previously described[122]. In brief, DIAMOND[265] (v0.9.14) BLASTp[101] (v2.6.0) was used to identify the most likely taxonomic affiliation of a sequence.

Additional datasets and benchmarking

Additional publicly available datasets were used to assess the performance of vRhyme under different scenarios and conditions. To assess binning of related types of viruses within the same sample, a total of 101 publicly available crAssphage sequences[147] were dereplicated using vRhyme (--derep_id 0.97 --frac 0.70 --method longest) to 86 non-redundant scaffolds. The non-redundant scaffolds were randomly fragmented as described previously into 791 fragments. To assess binning of megaphages and eukaryotic viruses with large genomes, the 540 kb Prevotella phage Lak C1[179] was randomly fragmented into 51 fragments, and four different eukaryotic viruses[379,380] with genome lengths ranging from 154 kb to 201 kb were each randomly fragmented into 11 to 19 fragments. To assess binning of active and dormant prophages, VIBRANT was used to predict prophage regions for 10 active prophages from 3 different hosts and 24 dormant prophages from 5 different hosts. Activity or dormancy was determined according to respective studies described elsewhere[339,343,344] and validated using PropagAtE[381] (v1.1.0). Whole prophage scaffolds from the same host genome were binned together. Details on the studies, reads used, scaffolds, and accession numbers can be found in **Supplementary Table 1**.

To validate protein redundancy, NCBI databases (RefSeq and Genbank, release July 2019) were queried for “prokaryotic virus” as before and genomes greater than 3 kb in length were retained. Likewise, NCBI databases (RefSeq and Genbank, release September 2021) were queried for “eukaryotic virus” and genomes greater than 20 kb in length were retained. Proteins were predicted using Prodigal (-p meta) for 15,238 prokaryotic and 557 eukaryotic viruses. Protein redundancy was calculated per genome based on the method described for vRhyme, with the exception that proteins could be redundant if encoded along the same scaffold.

Effect of number of samples

The effect of the number of input samples on vRhyme performance was done by stepwise increasing the number of BAM files used to calculate coverage from one to the maximum number of samples for a given dataset. To do this, samples were arranged in descending order, starting at the sample with the greatest total coverage across all scaffolds and were stepwise combined, ending with the sample with the lowest coverage.

Visualizations

All plots and visualizations were done using Matplotlib[228] (v3.2.0) and Seaborn[382] (v0.11.0). Genome alignment visualizations were made using EasyFig[227] (v2.2.2) and Geneious Prime 2019.0.3. Genome alignments to identify percent sequence identity were made using progressiveMauve[358] (development snapshot 2015-02-25). vConTACT2[266] (v0.9.19, --rel-mode Diamond --db 'None' --pcs-mode MCL --vcs-mode ClusterONE, ClusterONE[383] v1.0) was used to construct protein clustering networks and visualized using Cytoscape[267] (v3.7.2).

Results

vRhyme overview and workflow

The vRhyme workflow is done in five steps: read coverage processing, sequence feature extraction, supervised machine learning, iterative network clustering, and bin scoring (**Figure 1**). The base input to vRhyme are the assembled scaffolds or contigs to be binned (hereafter scaffolds) with a set minimum size of 2 kb. For optimal results, only virome scaffolds or predicted virus scaffolds should be used as input, though vRhyme can function with the input of an entire metagenome. An initial dereplication step to remove redundant input scaffolds is optional. Next, scaffolds are compared pairwise by read coverage composition per sample, which is a proxy for

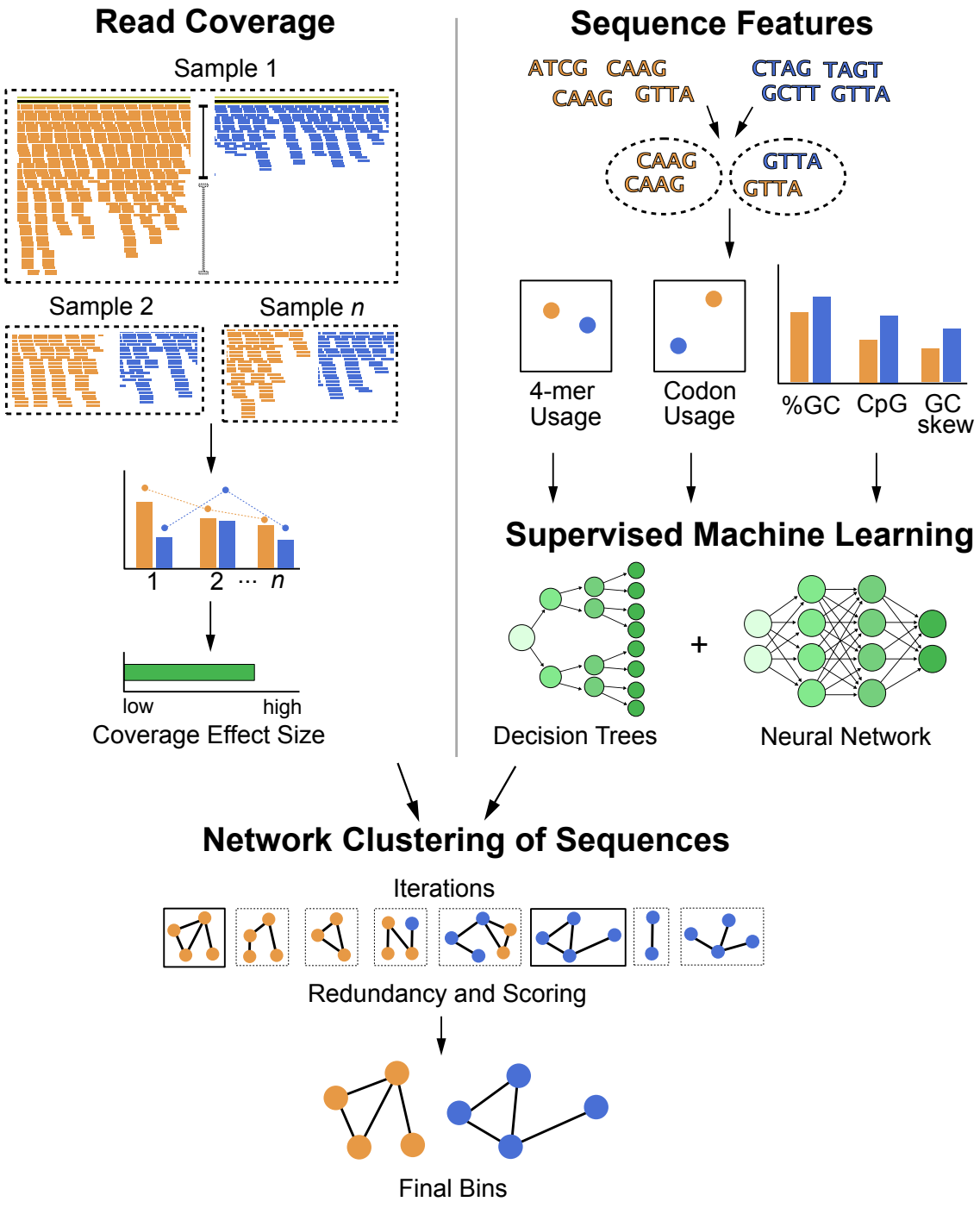


Figure 1. Flowchart of vRhyme workflow and methodology. Scaffolds are compared pairwise by read coverage effect size differences using single or multiple samples (top-left), followed by sequence feature distance comparisons (top-right). Multiple iterations of network clustering of putative bins are generated with edge weights representing normalized coverage effect size and supervised machine learning probabilities of sequence feature similarity (center). The bins are refined by KMeans clustering, and the best set of bins from a single iteration are identified after identifying protein redundancy and scoring (bottom).

relative abundance. vRhyme performs optimally with an input of multiple samples (i.e., coverage files) for more robust coverage co-occurrence estimations, but it will function with a single sample input with a minor decrease in performance. Statistically dissimilar scaffolds by coverage composition are screened out and the remaining potential pairs are compared by nucleotide feature similarity. Seven total nucleotide and gene features are used to classify pairs as similar versus dissimilar using two supervised machine learning models (decision trees and neural network). Following this step, potential connections are made between scaffolds based on similarity in read coverage and nucleotide features. These connections are used to create weighted networks that are further refined into genome bins using KMeans clustering. The entire process of read coverage comparison, nucleotide feature machine learning and weighted network refinement is performed over several *binning iterations* in parallel. vRhyme has 15 built-in presets of thresholds for Cohen's *d*, machine learning model probabilities, and network edge weights. The number of presets used is equivalent to the number of binning iterations completed. A list of all presets and their hierarchy can be found in **Supplementary Table 2**. Each bin within all binning iterations is scored according to protein redundancy, a proxy for contamination, and the best binning iteration by sequences binned, bins generated, and redundancy metrics is selected. The bins within this best binning iteration are reported along with relevant metadata, including number of members and total protein redundancy. Alternative binning iterations are likewise saved if manual inspection and selection of a different iteration is desired.

Assessment of binning quality

To evaluate vRhyme, we first benchmarked vRhyme against reference datasets and compared the performance to several available binning tools, all of which are built for microbes.

Many binning tools and wrapper software were not suitable for viral binning due to reliance on microbial single copy genes. We were able to successfully compare vRhyme to MetaBat2[92], VAMB[95], CoCoNet[118], CONCOCT[94], and BinSanity[367] on nine datasets curated from metagenomic data (see Methods). The nine datasets were comprised of 999 non-redundant and putatively complete viral genomes that were split into 4,324 sequence fragments of varying lengths between 2 kb and 20 kb. Of these, 1,118 fragments were less than 5 kb, 1,361 were greater than 5 kb and less than 10 kb, and the remaining 1,854 were greater than 10 kb. The average length was 9.4 kb. Although these fragments were derived from datasets not represented in the machine learning training dataset, we first verified that the fragments were distinct and would not result in a bias associated with an overfitted machine learning model. Based on BLASTn similarity at 70% identity and 60% overlap, only 255 (~6%) of the 4,324 fragments were represented in the machine learning model training dataset, with all but two of the represented fragments being from the same human gut dataset.

A total of 17 different evaluation metrics were used, including five traditional metrics for recall, precision, accuracy, specificity, and F1 score (**Figure 2**). The five traditional metrics were calculated according to the true positive, true negative, false positive, and false negative rates of binning fragments together from the same or different source genomes (**Supplementary Table 3a**). Note that the machine learning models were not benchmarked individually since performance is measured based on the entire pipeline. vRhyme yielded the highest F1 score, the harmonic average of precision and recall, with an average of 0.87 across all nine datasets. MetaBat2 and VAMB performed equally with F1 scores of 0.81 and 0.82, respectively, but importantly VAMB only successfully binned three of the nine datasets due to input size requirements. vRhyme likewise yielded the highest, or equal to highest, average precision (0.94), accuracy (0.90), and specificity

Figure 2. Benchmarking performance metrics of vRhyme compared to MetaBat2, VAMB, CoCoNet, CONCOCT, and BinSanity. Each boxplot represents the results of 9 different datasets, except for VAMB in which three datasets are shown. In total, 999 non-redundant genomes artificially split into 4,324 sequence fragments are shown. For some plots, a dotted line is shown at 1.0 to indicate optimal performance. CONCOCT and BinSanity are partially shown on the Genome-to-Bin Ratio plot for better visualization; each yielded an average ratio greater than 2.0.

(0.96) compared to all benchmarked tools. Compared to MetaBat2, VAMB and CoCoNet, vRhyme likewise yielded the greatest average recall (0.80). CONCOCT and BinSanity yielded the greatest average recall values (0.96 and 0.91, respectively) but at the expense of precision (0.45 and 0.44, respectively). At least for viral genomes, CONCOCT and BinSanity were found to not be suitable binning options. VAMB had suitable performance on the three datasets with enough input sequences, but VAMB is likely not an option for many applications of binning viral genomes due to requiring many input sequences (e.g., tens of thousands[95]) for optimal performance. Based on these metrics, vRhyme performed exceptionally in binning viral genomes but did not considerably improve on the performance of MetaBat2.

The remaining 12 evaluation metrics were calculated according to complete genomes and individual bins. These included evaluating if genomes were placed into a single or separate bins, and if bins contained fragments from a single or multiple source genomes. These metrics were better able to show the distinct performance of vRhyme compared to the other tools (**Supplementary Table 3b**). Namely, vRhyme was better able to reduce the following: placement of genomes into separate bins, placement of fragments from multiple source genomes into a single bin, and binning circular scaffolds representing entire genomes. Importantly, this was not at the cost of reduced fragment recall by vRhyme. To combine these metrics, we created a genome score and bin score that considered recall and precision as a substitution for F1 score. For genome scores and bin scores, respectively, vRhyme (0.89 and 0.96) outperformed, or was equivalent to, MetaBat2 (0.77 and 0.93) and VAMB (0.90 and 0.93). Again, it is important to note that VAMB only successfully binned three of the nine datasets. For CoCoNet, CONCOCT, BinSanity, genome scores (0.66, 0.74 and 0.70, respectively) and bin scores (0.65, 0.48 and 0.18, respectively) reflected the propensity to “over bin” distinct genomes together into one bin. CoCoNet did not bin

any sequence in two of the datasets, and after removal of these zero-values, the average genome score and bin score for CoCoNet both increased to 0.84.

Furthermore, we evaluated how well vRhyme bins compare to the input, unfragmented genomes. First, using CheckV[370] we show a distinct change in genome completeness estimation in the binned versus unbinned sequence fragments. vRhyme was able to recapitulate the

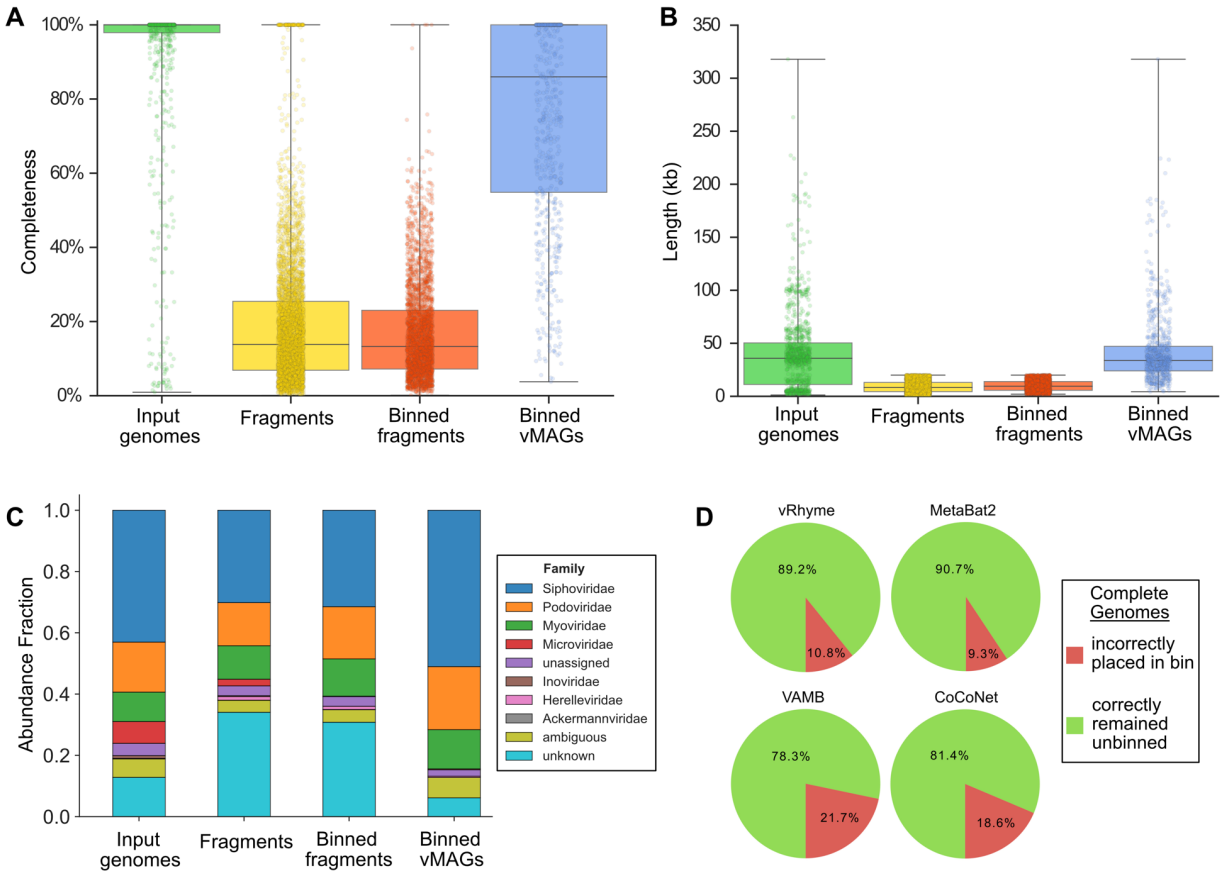


Figure 3. Impact of binning with vRhyme on the benchmarking datasets. For **a-c**, the putatively complete unsplit input genomes, generated sequence fragments, binning sequence fragments, and vRhyme bins (vMAGs) are compared. (**a**) Estimation of genome completeness using CheckV. (**b**) Sequence or vMAG nucleotide length. For **a-b**, each dot represents a single sequence or vMAG. (**c**) Estimation of taxonomy at the family level using a custom analysis script. "unassigned" represents a taxonomic classification to a group with an unassigned family, "ambiguous" represents equal assignment to multiple families (typically Caudoviricetes), and "unknown" represents the inability to make a prediction. (**d**) Evaluation of vRhyme, MetaBat2, VAMB, and CoCoNet for the binning of complete genomes. The expectation is that complete genomes should remain unbinned as uncultivated virus genomes (UViGs).

completeness of the input genomes (**Figure 3a**). This is supported by a similar observation in the length of the input genomes versus the bins (**Figure 3b**). Moreover, we estimated the taxonomy of the input genomes, fragments, and binned vMAGs. We identified a distinct decrease in the ability to identify taxonomy of the fragments, which were rescued by binning (**Figure 3c**). The identifiable difference in the vMAGs is a lack of Microviridae. Yet, this is to be expected since the small genome size of Microviridae (<10 kb) typically results in near-complete scaffolds that appropriately remain unbinned. Finally, we evaluated whether vRhyme could distinguish the source scaffolds. To do this, each of the nine datasets were binned, but the scaffolds were not fragmented. The expected result is that none of the circular scaffolds should bin together. Although vRhyme did bin ~11% of the whole scaffolds, it was a marked improvement on VAMB and CoCoNet (**Figure 3d**).

Benchmarking vRhyme on marine viromes

We next applied vRhyme to the Global Ocean Virome 2 (GOV2) database[10] and compared the results to MetaBat2 and CoCoNet. For metagenomic datasets such as GOV2 the expected number of scaffolds to bin and the number of bins is unknown. First, all scaffolds from the GOV2 database were limited to scaffolds at least 5 kb in length and dereplicated by 98% identity. Of the 108,947 input scaffolds, vRhyme binned 56,642 scaffolds into 13,175 bins, MetaBat2 binned 57,800 scaffolds into 11,826 bins, and CoCoNet binned 91,842 scaffolds into 9,914 bins. Despite the number of bins generated being relatively similar, the number of scaffolds binned was quite different. However, vRhyme yielded 15,106 redundant proteins whereas MetaBat2 (29,334) and CoCoNet (71,364) yielded more, indicating that vRhyme was likely more precise and generated

fewer contaminated bins (**Figure 4a**). In support of this, vRhyme generated 1,266 bins with 2 or more redundant proteins whereas MetaBat2 (1,648) and CoCoNet (2,743) generated more. When these likely contaminated bins were removed, vRhyme binned 48,251 scaffolds into 11,909 bins, MetaBat2 binned 33,351 scaffolds into 10,178 bins, and CoCoNet binned 35,380 scaffolds into 7,171 bins (**Figure 4b**). Based on protein redundancy, vRhyme was capable of binning far more viral scaffolds and generating more bins of low contamination compared to MetaBat2 and CoCoNet. Note, we identified bins with "low contamination" to be 0-1 redundant proteins based

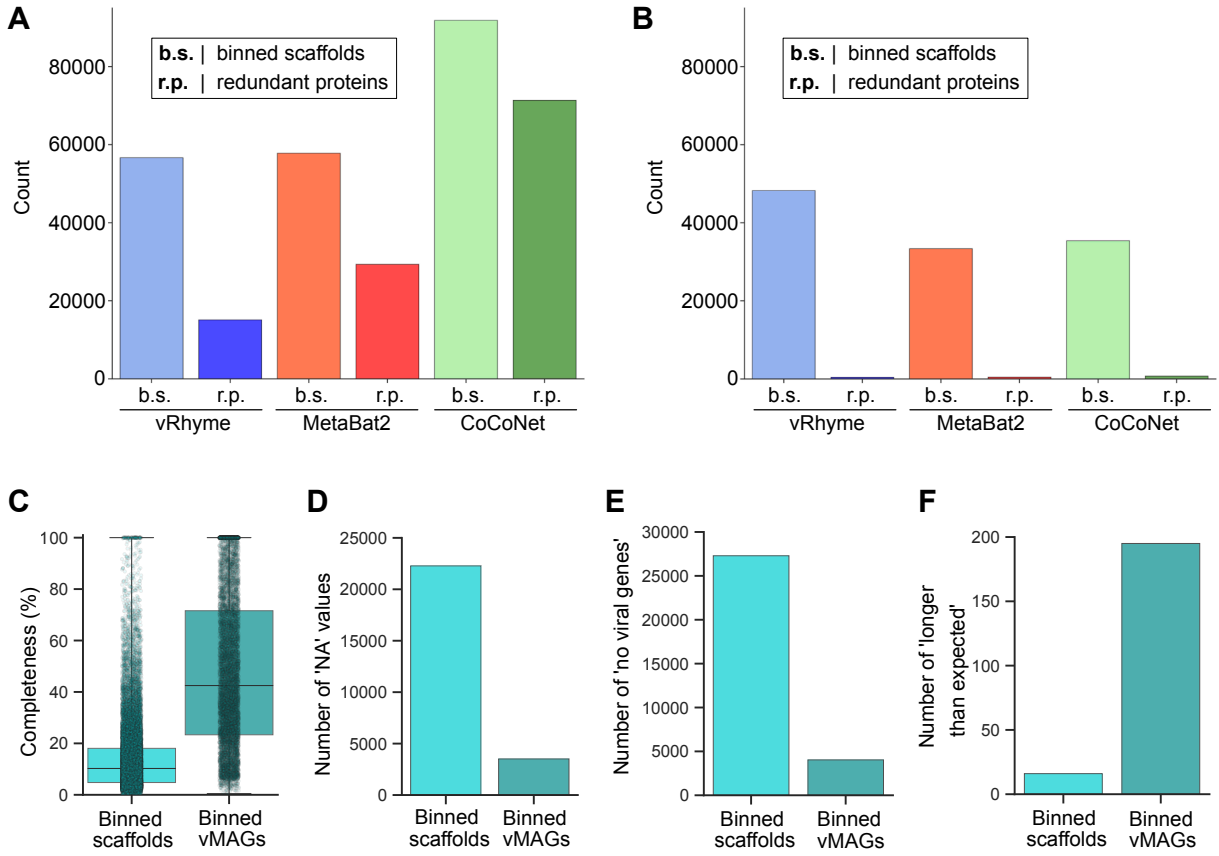


Figure 4. Benchmark binning and genome completeness evaluation of GOV2. Comparison of vRhyme, MetaBat2, and CoCoNet (**a**) raw results and (**b**) low contamination filtering results by the number of scaffolds binned and identified redundancy. For vRhyme only, CheckV was used to identify (**c**) the estimated completeness values, (**d**) number of 'NA' completeness values, (**e**) number of 'no viral genes' scaffolds/vMAGs, and (**f**) number of 'longer than expected' scaffolds/vMAGs for the low contamination results of individual binned scaffolds as well as vMAGs.

on a benchmark of prokaryotic and eukaryotic viral genomes from NCBI databases (**Supplementary Figure 2**). Contamination was not estimated using CheckV as that metric does not consider contamination of multiple viral genomes, but rather contamination of non-viral sequences.

We also estimated the completeness of the 11,909 low contamination vRhyme bins and the individual 48,251 scaffolds that generated those bins using CheckV. The binned scaffolds individually yielded 25,969 (53.8%) completeness values with an average of 14% complete, 79 estimated to be 100% complete, 22,282 (46.2%) with 'NA' completeness, and 27,295 (56.6%) with 'no viral genes detected'. The scaffolds within each bin, after being linked into vMAGs, yielded 8,393 (70.5%) completeness values with an average of 48% complete, 775 estimated to be 100% complete, 3,516 (29.5%) with 'NA' completeness, and 4,039 (33.9%) with 'no viral genes detected' (**Figure 4c-e**). There was an increase in the number of vMAGs (195, 1.6%) versus individual scaffolds (16, 0.03%) that were estimated to be 'longer than expected', potentially due to a marginal rate of multiple genomes being binned into a single vMAG (**Figure 4f**). Overall, vRhyme generated vMAGs with greater average completeness to aid in downstream analyses and interpretations, even with high complexity or large datasets such as GOV2.

Discovery of vMAGs in human skin metagenomes

To demonstrate the ability of vRhyme to aid metagenome analyses and discovery, we applied vRhyme to 270 human skin metagenomes[376]. Viruses were predicted from a cohort of 34 individuals with eight body sites (*Af, Al, Ba, Na, Oc, Tw, Um,* and *Vf*) sampled per individual (see Methods). From all individuals, 10,601 viral scaffolds were identified and binned, across eight different body sites individually, into a total of 849 vMAGs representing 2,794 viral scaffolds.

Although bins with redundant proteins may in fact be a single genome or partially redundant copies of a single genome, we ignored all vMAGs with greater than one redundant protein for analysis to yield 762 vMAGs representing 2,413 viral scaffolds, leaving the remaining 8,188 as discrete viral scaffolds (**Supplementary Table 4**) (**Figure 5a**). The taxonomic classification of UViGs pre-binning, UViGs and low redundancy vMAGs post-binning, and vMAGs-only displayed that most bins were constructed of genomes from the class Caudoviricetes, similar to the observed taxonomy pre-binning (**Supplementary Figure 3**). The bins were comprised of an average of 3.2 scaffolds each. In total we identified seven bins, representing separate body sites, that were present across at least 30 individuals (**Figure 5b**). In addition, two bins of unique characteristics were identified and examined in detail.

The first such bin contained 22 members (Tw bin 8), more than what would be expected for a viral bin, and aligned to a reference Herelleviridae phage (Staphylococcus phage phiSA_BS2) (**Figure 5c**). Herelleviridae infecting abundant *Staphylococcus* on the skin are likely to be highly relevant to skin ecology and disease[384]. Before binning, each of the 22 members were identified by CheckV as low-quality genome fragments with individual completeness estimations ranging from 1.8% to 7.1%. The fragments averaged 5.2 kb in length and ranged from 2.6 kb to 10.0 kb. After binning, the final bin was 115 kb in length and identified as a high-quality genome with 100% completeness by CheckV. The reference phage genome is 143 kb, suggesting the true completeness of the bin is likely 80% to 100%. All CheckV results for the skin metagenomes can be found in **Supplementary Table 5**.

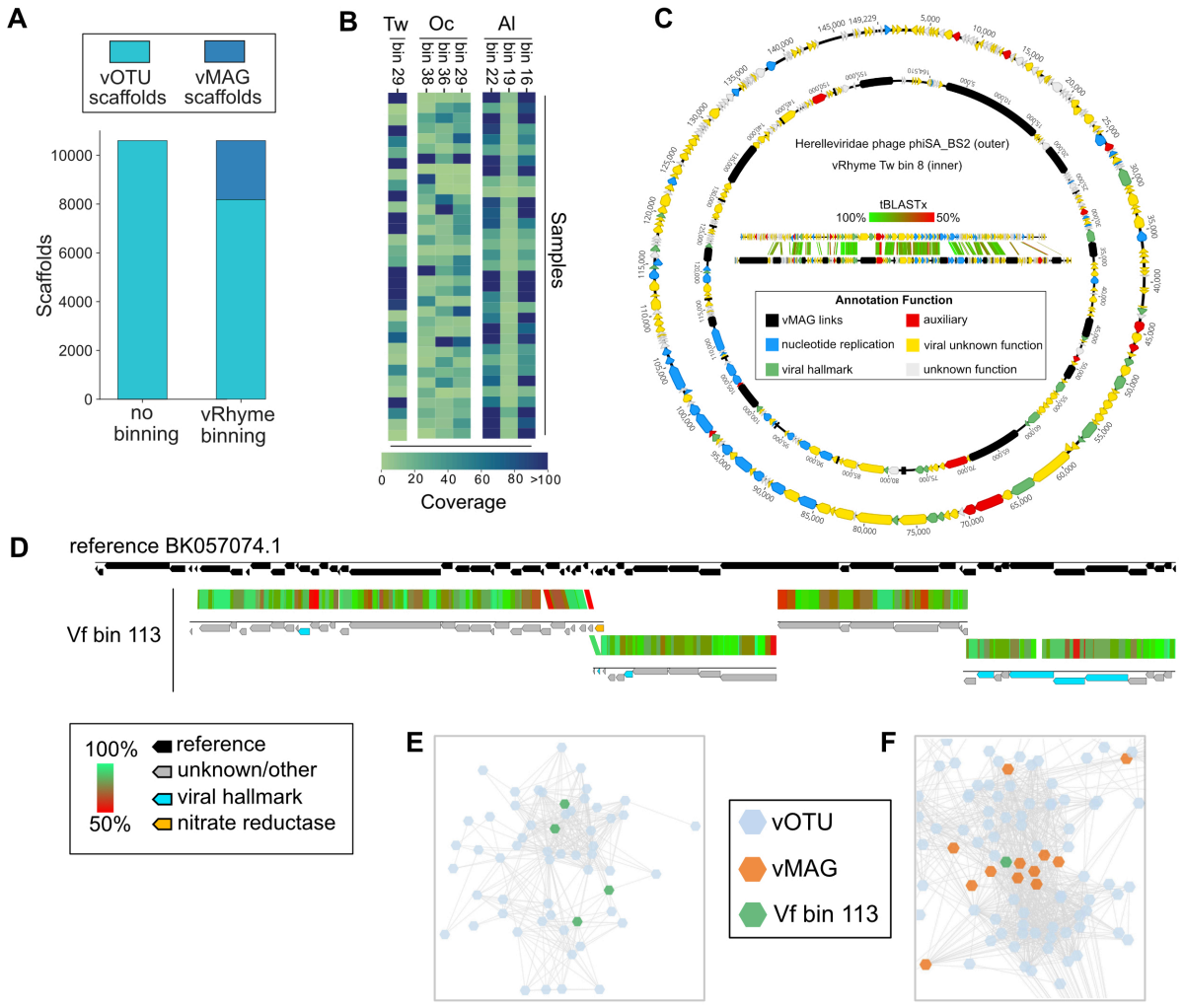


Figure 5. Binning improves and expands the analysis of viruses from human skin. (**a**) Comparison of the number of original viral scaffolds identified across all individuals before and after binning. (**b**) Heatmap of coverage for the seven common bins per individual. (**c**) Genome visualization and alignment of Herelleviridae reference phiSA_BS2 (outer) and Tw bin 8 (inner). Each arrow represents a predicted open reading frame and black bars are artificial connections between vMAG scaffolds. (**d**) Alignment of vRhyme Vf bin 113 to the closest reference virus Siphoviridae isolate ctiXA4 (BK057074.1). Each of the four scaffolds were independently aligned by tBLASTx similarity. The *narG* AMG is labeled in yellow and viral hallmark annotations are labeled in light blue. (**e**) Representative cluster from all input viral scaffolds generated by vConTACT2, with the four Vf bin 113 scaffolds labeled in green. There are no connections between any of the four green scaffolds. Each dot represents a single scaffold. (**f**) Partial network from all vRhyme binned and unbinned viral scaffolds generated by vConTACT2, with vMAG bins labeled in orange and Vf bin 113 in green. For **e**,**f** Complete network diagrams can be found in **Supplementary Figures 4 and 5**.

The second bin of interest contained 4 members (Vf bin 113), with one encoding a nitrate reductase (*narG*) auxiliary metabolic gene (AMG) (**Figure 5d**). The *narG* was positioned as the last gene on a scaffold, and conventional approaches for AMG validation would suggest discarding the AMG as likely bacterial contamination. However, binning aided in the validation of the AMG as likely to be correct. The first line of evidence was the lack of any integrase or lysogenic viral signatures on any of the four binned scaffolds, suggesting the AMG is not from bacterial contamination resulting from host integration. Second, alignment of all four scaffolds to the nearest reference genome (Siphoviridae isolate ctiXA4) displayed that the AMG was situated at the intersection of two scaffolds within the genome rather than at a genome end. CheckV identified each member as low-quality with completeness values of 11.6% to 28.0% for the respective 7.4 kb to 16.8 kb scaffolds. The bin was estimated to be of medium-quality with a completeness of 74.9%, or 92% based on the length of the closest reference genome. Moreover, one of the four scaffolds lacked characteristic viral annotations to aid with manual inspection or analyses such as phylogeny, yet binning with the other scaffolds containing viral hallmark and nucleotide replication annotations was able to validate the scaffold as viral and place it in better genomic context for analysis. Therefore, binning was able to not only generate a more complete sequence, but also validate the presence of an understudied and ecologically important AMG. Using vConTACT2[266], we clustered all of the individual, unprocessed viral scaffolds (**Figure 5e**) in addition to the bin with the complete binning results (low-contamination bins plus unbinned scaffolds) (**Figure 5f**). Clustering of the individual scaffolds placed all four scaffolds of the bin into a single cluster distinct from other groups, yet as anticipated none of the scaffolds of the bin were connected. Clustering of the binning results yielded more connections between scaffolds and

vMAGs and better placed the bin within evolutionary and community relationship contexts. Complete vConTACT2 networks can be found in **Supplementary Figures 4 and 5**.

Discussion

Binning viral scaffolds into vMAGs is uncommon, with most or all remaining as discrete virus operational taxonomic units (vOTUs) or uncultivated virus genomes (UViGs)[308]. We believe adopting a more genome-centric approach for UViGs will enable innovative discoveries, such as the construction of large or highly heterogenous viral genomes that often assemble into dissimilar fragments. Here, we have presented vRhyme and demonstrate that the "one scaffold, one virus" convention can skew interpretations of a virosphere and the interactions of its viral community members. To address this, vRhyme enables the binning of viral genomes into vMAGs using a virus-centric approach, unique from existing binning software, in an easy to use and reproducible command line tool.

In addition to performance benchmarks on artificial and real metagenomes, we evaluated the robustness of vRhyme by binning artificially fragmented NCLDV, megaphage, large eukaryotic virus, crAssphage, active and inactive integrated prophage, and microbial genomes (**Supplementary Information, Supplementary Table 6**). vRhyme was largely capable of precisely binning these unique and complex viral datasets. However, notable exceptions were difficulties with separating multiple inactive (non-replicating) prophages from the same host genome as well as binning non-viral genomes, though the latter was an anticipated limitation. Moreover, we displayed that vRhyme is efficient and likely precise in binning large and complex datasets using GOV2 and agricultural soil viromes[106] (**Supplementary Information, Supplementary Table 7**). In total, we hope that with the availability of vRhyme as a reliable

binning tool, vMAG construction will become a common practice and adopted into existing frameworks of studying viral ecology, host associations, community interactions, evolution, and biogeochemical cycling.

To further evaluate the computational capabilities of vRhyme or potential restraints, we assessed the effect of the coverage calculation methods, the number of input coverage samples and the effect of user-modifiable parameters on performance, as well as the runtime, memory usage and reproducibility of binning (**Supplementary Information**). We found that vRhyme performs optimally with multiple input samples for more robust coverage variance comparisons, though the optimal value depends on how the dataset or metagenome was constructed (**Supplementary Table 8, Supplementary Figure 6**). For example, a metagenome assembled from a single, standalone sample may perform suitably. As for modifying parameters, vRhyme likely will yield optimal results with the default settings due to the coverage calculation method employed and built-in binning iterations (**Supplementary Table 9, Supplementary Figure 7**). Furthermore, the runtime of vRhyme for average sized viral datasets was on the scale of seconds. The GOV2 dataset, the largest dataset evaluated, finished in 93 minutes with 2.3 GB of memory using 15 CPU threads (**Supplementary Table 10, Supplementary Figure 8**). Lastly, the methods employed by vRhyme allow it to be fully reproducible. Overall, we found the necessary requirements to be relatively low and even possible on personal laptop systems.

There are several important considerations in the binning of vMAGs that are unique from microbial MAGs. First, any viral scaffold not contained within a bin (vMAG) should be considered as a vOTU or UViG. This aligns with the "one scaffold, one virus" convention which is likely true for many viral genomes, especially circular and complete genomes. In the skin datasets presented here, ~23% of the viral scaffolds were binned into low contamination vMAGs and the remaining

~77% should still be utilized in analyses as discrete scaffolds. Second, an entire metagenome can be used as input to vRhyme, or viral binning in general, with the caveat that contamination of bins with non-viral sequences may be higher with the added advantage that fewer viral scaffolds may be missed. For example, many phage genomes are arranged in cassettes such that structural, nucleotide replication, lysis and auxiliary genes form distinct regions. If these regions were to assemble into separate scaffolds, virus identification may only identify a portion of the scaffolds, such as missing an auxiliary region, whereas binning may place them all together into a single vMAG. When applied to a synthetic dataset of predominately non-viral sequences, MetaBat2 performed better than vRhyme (**Supplementary Information, Supplementary Table 11**). Third, accurate read coverage profiles are crucial for accurate binning. This is true for all binning software that depend on differential coverage and is especially true for distinguishing bins of integrated prophages from a single host population. vMAGs representing prophages generated by vRhyme will likely represent the greatest fraction of redundant, contaminated bins.

Data Availability

vRhyme and all auxiliary scripts are freely available as open-source Python code at https://github.com/AnantharamanLab/vRhyme.

Acknowledgements

We thank members of the Anantharaman laboratory at the University of Wisconsin-Madison for helpful feedback and discussions. This research was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM143024 to K.A. and award numbers R35GM137828 and U19AI142720 to L.K. A.A. was funded by a University

of Wisconsin-Madison CIBM postdoctoral traineeship from the National Library of Medicine (T15LM007359). K.K. was supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison, and a William H. Peterson Fellowship Award from the Department of Bacteriology, University of Wisconsin-Madison.

Author contributions

K.K. and K.A. designed the study. K.K., A.A., and R.S. developed code and conducted bioinformatic analyses. K.K. and K.A drafted the manuscript. All authors (K.K., A.A., R.S., L.K., and K.A.) reviewed the results and approved the manuscript.

Chapter 7: Virus Genomics: What is Being Overlooked?

Kristopher Kieft[1,2] and Karthik Anantharaman[1]

[1]Department of Bacteriology, University of Wisconsin–Madison, Madison, WI, USA
[2]Microbiology Doctoral Training Program, University of Wisconsin–Madison, Madison, WI, USA

Publication:

Kieft, K. & Anantharaman, K. Virus genomics: what is being overlooked? *Current Opinion in Virology* **53**, 101200 (2022).

Abstract

Viruses are diverse biological entities that influence all life. Even with limited genome sizes, viruses can manipulate, drive, steal from, and kill their hosts. The field of virus genomics, using sequencing data to understand viral capabilities, has seen significant innovations in recent years. However, with advancements in metagenomic sequencing and related technologies, the bottleneck to discovering and employing the virosphere has become the analysis of genomes rather than generation. With metagenomics rapidly expanding available data, vital components of virus genomes and features are being overlooked, with the issue compounded by lagging databases and bioinformatics methods. Despite the field moving in a positive direction, there are noteworthy points to keep in mind, from how software-based virus genome predictions are interpreted to what information is overlooked by current standards. In this review, we discuss conventions and ideologies that likely need to be revised while continuing forward in the study of virus genomics.

Introduction

Genomics approaches for the study of viruses (infecting eukarya and archaea) and bacteriophages (phage; viruses infecting bacteria) has taken off in the last few years, much in part due to our ability to understand and interpret viral genomes from metagenomes. In fact, it is common to find a publication describing environmental virus genomics from the last few years that indicate viruses as the most abundant and diverse biological entities on the planet. As a scientific community, we are recognizing the extensive footprint viruses leave on all environments where life exists. For example, examining viral genomes has allowed us to discover metabolic genes encoded by viruses such as for photosynthesis and sulfur oxidation, and extrapolate the impacts of virus-directed metabolism on various biogeochemical processes [35,73,74,81,84,122,140,351].

Investigating viral genomes has also aided in the innovation of novel CRISPR-based genome editing technologies [22–24], further development of phage therapy applications [29,385], broader understanding of human gut dysbiosis [11–13], and more.

Unseen to our daily lives, viruses and phages are constantly modifying the planet around us through manipulation and/or lysis of their hosts [70]. Unfortunately, only a small fraction of all viruses that are estimated to exist have been cultivated in the laboratory. This has led to great interest in utilizing next-generation sequencing and metagenomics specifically, to catalog, explore, describe, and understand the diversity of viral genomes [10,308,386,387]. Through metagenomic methods and technologies, thousands of viral genomes can be acquired from a single mixed metagenome (mixed community) or virome (virus-specific) sample.

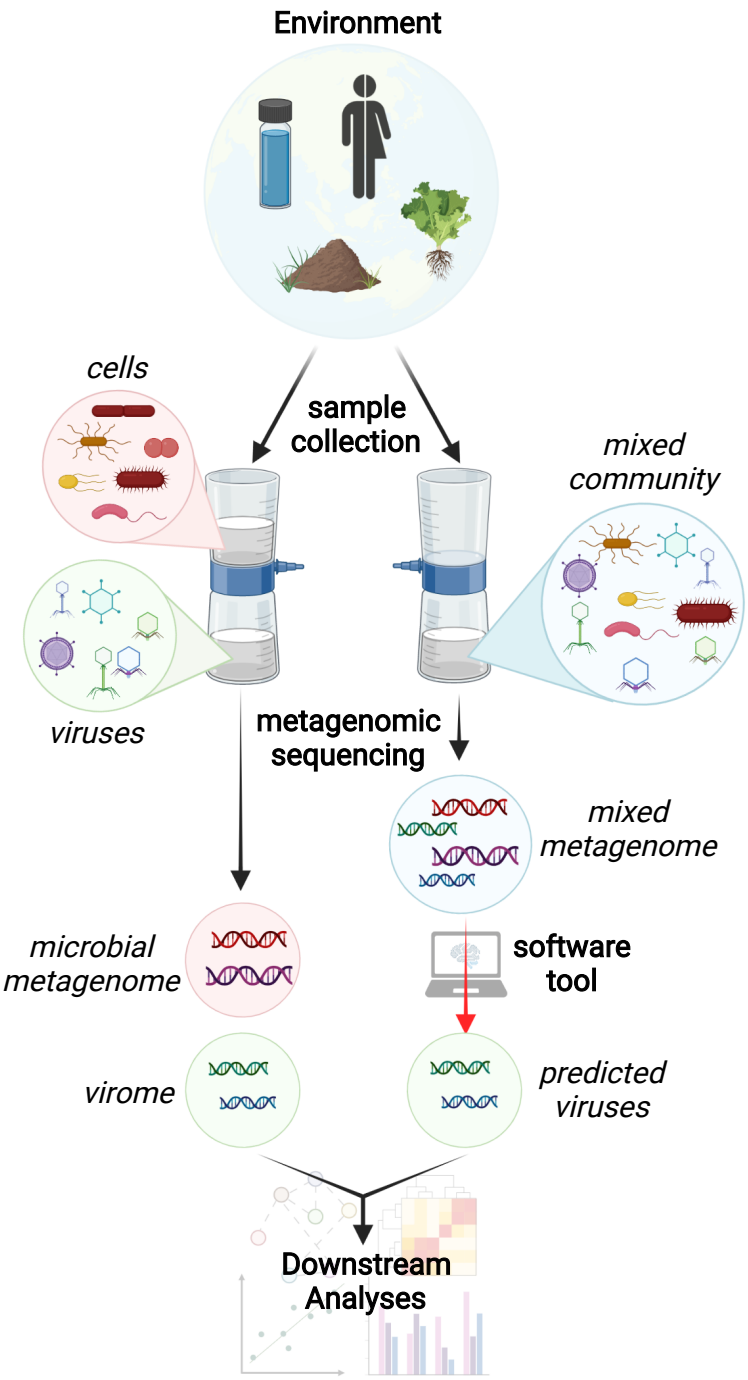


Figure 1. Sample collection and metagenomic sequencing of viruses. Virus genomes can be identified by physical separation from cells (left) or by software tool prediction (right) preceding downstream analyses.

There are two general methods by which to obtain genomic information to study viruses using metagenomics: extraction and sequencing of viromes, and virus prediction from mixed microbial metagenomes (Figure 1). A virome differs from a conventional mixed microbial metagenome in that it is the physical separation, collection, and sequencing of virus-like particles (VLPs) from a sample. Methodologies of VLP collection vary considerably and require modification depending on the source environment (e.g., soil, aquatic, human gut). Each method comes with its own use-case utilities, biases, and ease-of-use, and no one method is globally accepted in the field. A virome can be described as an *in situ* method of virus discovery. On the other hand, virus prediction is the *in silico* discovery of virus sequences from a metagenome, or even a virome; a software tool or manual sequence inspection is used to separate viral from non-viral sequences within a mixed community. Notably, there are distinct differences between these two methods that impact the way in which the data is analyzed. For studies specifically focused on the viral fraction of an ecosystem, VLP sequencing of the virome can yield results best suited for studying viral communities [106]. Virome samples are often better at capturing low abundance viruses but may exclude viral genomes that are in an intracellular state (e.g., non-replicating proviruses and virocells) [70]. Conversely, predicting viral sequences from bulk metagenomes can provide context of the viruses and microbes together within the same sample, such as allowing for more accurate host predictions or identifying intracellular viral genomes [381,388].

In the last few years there has been a rapid expansion in the knowledge of viruses on a global genomics level by using metagenomes. Here, we slow down and take a step back to ask what is being overlooked? Considering the current state of virus genomics, where should conventions be broken, and innovations be made? To do this, we will explore some of the methods available to extract viral sequences from metagenomes and describe best practices of how those

sequences can or should be analyzed. Here, we will focus on software-based virus prediction methods and their benefits, utilities, flaws, biases, and future directions.

Sweeping contamination under the rug: balancing recovery and false discovery

Virus prediction from mixed metagenomes is powerful in that it allows for an entire sample to have nucleotides extracted and sequenced while maintaining the integrity of the original microbial community comprised of organisms and viruses. A substantial number of software tools are currently available to predict viruses from nucleotides with varying methods, degrees of precision, and recovery capabilities [113,116,117,294–296,389–394]. In all cases, it is vital to consider the reality of these predictions in that all computational methods have drawbacks (Figure 2a, Table 1).

Virus prediction, for the vast majority of implementations, do not encompass all viruses in a sample due to loss in recovery, low sequencing depth of the viruses compared to microbes, or biases against certain viral families. Therefore, when using software to predict viral sequences, the recovered viruses will represent a subset of the true composition. These results can be influenced by the specific computational methods utilized by different tools or universal limitations in available methods [395]. For example, all currently available tools are limited by known virus diversity and struggle to predict viruses with entirely novel sequences. Many tools are also biased toward dsDNA viruses and phages due to dsDNA-centric databases and sequencing methods. Likewise, viral genome sequences comprised mostly of genes or features common to both viruses and organisms are difficult to identify accurately. These biases have the potential to leave behind viruses with novelty to reference databases or regions of recent recombination without close inspection [179,361]. In general, all software tools can only find viruses that appear similar to what we already know about due to reliance on reference-based prediction methods (see *the reference-free*

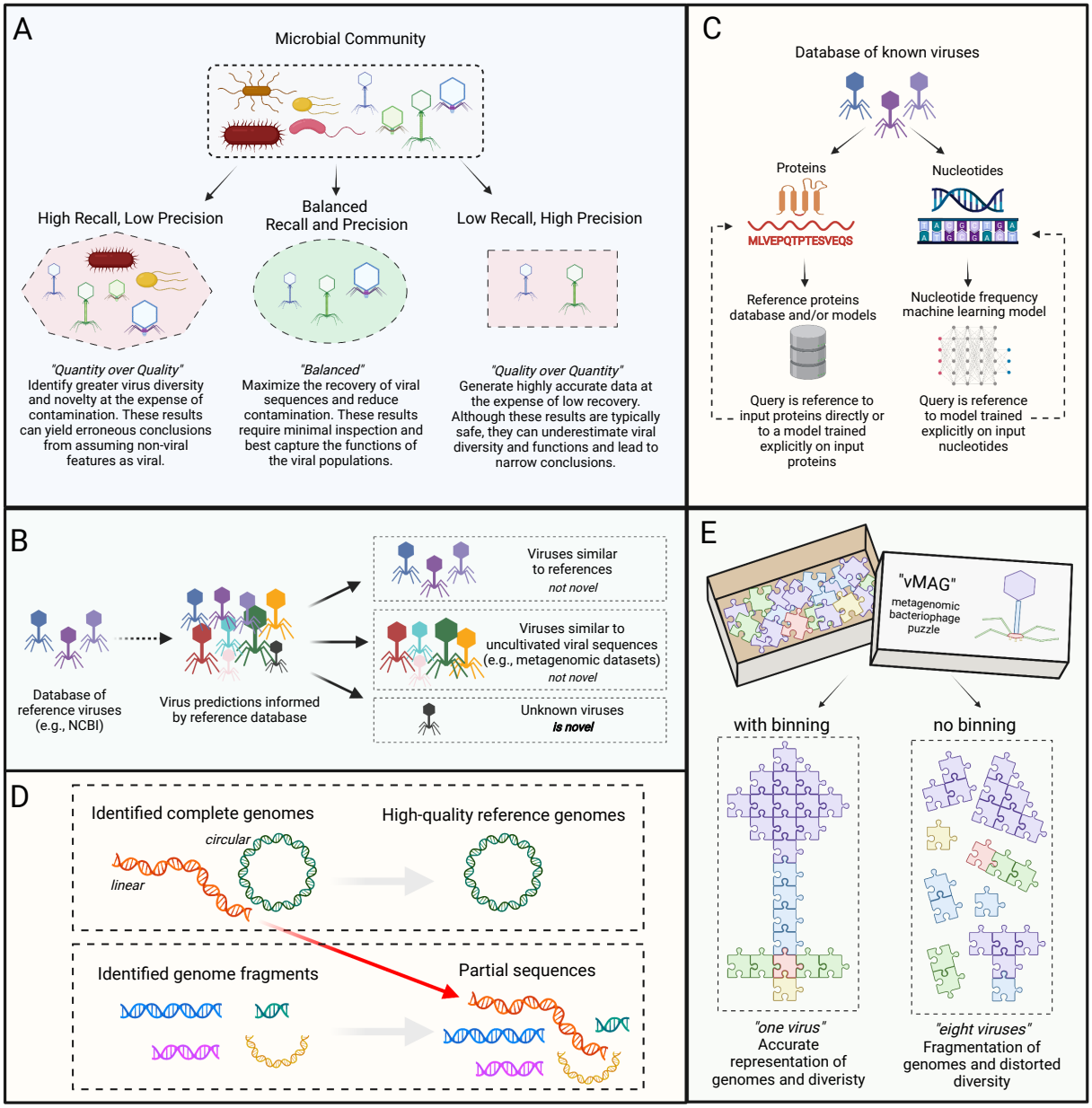


Figure 2. Conceptual summary diagram. **A**: comparison of general virus prediction strategies utilized by software tools, from variable recall and precision capabilities to a balanced approach. **B**: categorization of virus predictions as "not novel" or "novel" according to similarity to reference databases and datasets of uncultivated viral sequences. **C**: the reference-free fallacy; visualization of how virus prediction software tools, whether protein annotation-based (left) or nucleotide feature-based (right), are all inherently referenced-based. **D**: the fate of complete linear versus circular viral genomes in interpreting metagenomic data. **E**: illustration of a viral genome either binned into a vMAG (left) or analyzed as individual fragments (right); each sequence fragment is represented by puzzle pieces.

fallacy below). This limitation has been addressed by incorporating non-reference (e.g., metagenomic) sequences into software training algorithms, but with the caveat that contamination of virus predictions or virome extractions is not uncommon [9,116].

Contamination, or false discovery, of non-viral sequences is a feature of all virus prediction software and should not be ignored. That is, not all recovered sequences predicted to be viruses should be included haphazardly into analyses [396]. In most cases, the time, expertise, and/or computational resources are not available to manually validate all recovered viruses. However, the reality behind the precision of predictions should be made clear, such as providing details of how the prediction results may have been validated including software-specific cutoffs and identification of viral hallmark genes [397]. This is especially relevant when considering the ratio of recovery to precision. For example, reporting numbers of high virus identifications (high recovery) at the expense of the validity of those identifications (low precision) yields seemingly valuable but fundamentally flawed data. Low precision can result from the poor performance of a software tool, incorrect usage of a software tool (e.g., wrong implementation or retaining low probability or scored predictions), inclusion of many short sequence fragments (e.g., less than 3 kb), and other factors.

The following sections stem from the original biases and limitations of the current state of virus prediction. By exploring these topics, we aim to shed light on the potential advancements in computational methods or inconsistencies in interpretations for viral metagenomic data.

Of reference and reality

Many of the gold standards (i.e., trusted reference sequences) for viral genomes are deposited in public repositories such as NCBI databases [207,210]. These sequences are utilized by

various software tools beyond virus prediction, such as for prediction of hosts of viruses, prediction of virus taxonomy, functional annotation, genome quality assessment, and more [266,370,398]. However, this presents significant biases owing to the small and non-diverse composition of NCBI databases, relative to nature. The diversity of viruses by taxonomy and sequence composition within NCBI databases is estimated to be far less than what can be identified in nature and is primarily limited to viruses that have been cultivated on a limited number of hosts, mostly those of clinical significance or as a model research system [1]. Considering virus prediction software tools are reliant on these reference databases, it is clear that there are pitfalls associated with assuming that reference sequences fully mimic natural reality.

Similarly, the designation of viral genomes as "novel" according to a database search is not equivalent to true novelty. True novelty refers to if a given genome has yet to be identified by other sources and is not deposited in another database. For example, a search of NCBI databases excludes the majority of metagenome-derived viral sequences, many of which can be found throughout the literature and in curated databases [9,10,388,399]. Therefore, a virus may be novel with regard to reference database sequences, but not actually represent a truly novel sequence. Another source of novelty can be if the given sequence contains features yet to be discovered or broader implications that have yet to be identified. For example, the identification of crAssphages as highly abundant in the human gut came after representative sequences were deposited into databases [180] (Figure 2b, Table 1).

The reference-free fallacy: no such thing as a reference-free virus prediction

Many virus prediction software tools are based on *bona fide* genomes derived from NCBI RefSeq, which is mainly composed of isolated and cultivated viruses that serve as reference

systems. There are two broad categories of tools according to the methods used: nucleotide sequence features (e.g., VirFinder) and protein similarity (e.g., VIBRANT), or a hybrid of both (e.g., VirSorter2) [113,116,117]. For either category, machine learning has become a powerful approach for identifying patterns to increase prediction reliability and specificity [400]. However, this has led to some misconceptions to believe that "reference-free" refers to complete independence from reference databases, whereas "reference-based" refers to the use of protein annotation methods based on the annotations of reference viruses. Conversely, we advocate there is no tool completely reference-free and rather all tools are inherently reference-based in some manner (Figure 2c, Table 1).

For a tool that utilizes protein annotation, the reliance on reference sequences is in the form of prediction models built from a protein database [213,306,401], which is a clear reference-dependent method. Namely, only reference proteins are able to be annotated, queried, and subsequently analyzed. On the other hand, a tool that strictly uses sequence features (e.g., tetra-nucleotide frequency) does not necessarily need to rely on a database, but can rather rely on a machine learning model. This machine learning model can be perceived as reference-free, but similar to a protein database, the model too is dependent on the reference sequences used to train it. Therefore, for both categories of tools there is a direct reliance on reference sequences, making them both inherently reference based. A more accurate distinction would be "database-dependent" or "database-free" methods. Even manual verification of virus predictions is not reference-free as this method typically involves searching through protein annotations (e.g., phage structural hallmark proteins) and other reference-informed signatures (e.g., gene density and gene strand switch frequency) [402].

Moreover, it is important to note that the reference sequences used to compare, train and test software tools and/or machine learning models typically all come from the same genetic pool (i.e., NCBI databases). This perpetuates biases: biases against rare virus groups and biases in accurate comparisons. First, it is estimated that the true diversity of viruses in nature has yet to be captured by the sequences available on NCBI databases [208,308,399]. This results in a lack of representation of more rare viruses or simply those that have yet to be isolated/cultivated [43,179,403,404]. Since virus prediction tools are inherently reference-based, this leads to perpetual biases towards identifying viruses we already know about, with rare occasions of identifying a truly novel species [403]. Second, the utilization of NCBI databases for assessing available software tools results in an inherent loss of fair comparisons. It is becoming increasingly difficult to generate a comparison dataset of gold standard viral sequences that does not, in some capacity, represent the sequences used to train existing tools. This is due to the limited size of NCBI databases. Especially for tools that utilize machine learning, evaluating a tool with a sequence that was used to train that tool results in inflated, positive performance. The common work around is to only include viral sequences submitted to NCBI databases after the dates of publication for tools to compare, but this also results in biases, such as the inclusion of viruses nearly identical to those submitted previously. This latter example can be addressed by removing any identical sequences via dereplication, though this is seldom employed. In attempts to solve this issue and generate comprehensive, fair datasets for future software tool development and comparison, more focus and better curation standards need to be placed on the construction of reference sequence datasets.

Linear genomes can be complete: where did all the linear genomes go?

Identifying complete viral genomes from sequencing data allows for more robust analyses compared to fragmented, partial genomes. Automated methods to predict complete viral genomes focus on circularization signatures, namely the identification of terminal nucleotide repeats (direct or inverted) of free viral sequences or insertion sites of viruses integrated into their host's genome (proviruses) [116,117,295,296,370,392]. For free (lytic cycle) viruses, the identification of circularization can typically indicate with confidence that the given genome is complete. However, this method discounts complete linear genomes, such as those without identifiable terminal repeats [405].

Thus far, no high-throughput informatics method exists for the identification of complete linear genomes in the absence of circularization signatures [370,406]. This results in over-emphasizing circular genomes as the only gold standards in generating metagenomic-based reference genomes or the highest quality genomes in genomic datasets. Though these conclusions are not flawed on their own as correctly identified circular genomes are certainly of high quality, barring false positives [362], this overall bias against linear genomes has infiltrated the currently available literature (Figure 2d, Table 1). Speculatively, the ability to identify complete, linear virus genomes may allow for a more holistic view of a viral community or lead to novel discoveries of underappreciated viral groups.

Metagenomes are puzzles: an unfinished puzzle is still just pieces

Metagenomic assemblies reconstruct thousands to millions of sequence fragments (*contigs*) representing partial genomes, and rarely complete genomes. A common practice in the study of bacterial and archaeal genomes is to reconstruct metagenome-assembled genomes (MAGs) [309,407]. This is typically done through a method termed *binning* where anywhere from two to hundreds or even thousands of contigs may be grouped into a single, putative genome (*bin*).

When using short read (e.g., 75-300 bp) sequencing technology and assembly, many resulting contigs are less than 5 kb in length, with relatively few exceeding 20 kb. Consequently, bacterial and archaeal genomes that generally exceed 1,000 kb must be computationally binned into MAGs. Though long-read (e.g., 1-20 kb) technologies are advancing these boundaries, the construction of MAGs is typically still required. For bacteria and archaea, several software tools are available for binning and constructing MAGs [93,94,96,97,408,409].

Viral genomes range from as small as 3 kb to greater than 2,000 kb. Many identified phages are members of the class *Caudoviricetes* (formerly *Caudovirales*) which range considerably in size, but most are approximately 30 kb to 200 kb [108]. Interestingly, the convention accepted in descriptions of viruses derived from viromes or predicted from metagenomes is that a single contig represents an uncultivated viral genome (UViG) or virus population [308]. To assume each sequence represents a separate genome likely far overestimates viral diversity within a sample given the expected fragmentation of viral genomes. This is especially true for viruses that are rarer and would likely result in high genome fragmentation after assembly. The construction of viral metagenome-assembled genomes (vMAGs) would better represent the true composition of viruses within a sample. Importantly, UViGs still have utility in that any viral sequence left unbinned may represent an entire viral population, contrary to what is accepted for bacteria and archaea where unbinned sequences are typically discarded (Figure 2e, Table 1). This can be achieved by binning vMAGs using short- or long-read sequencing [410]. Despite this, few studies bin vMAGs, and those that do bin typically focus on viruses with the largest genomes [81,82,120,411]. This conspicuous discrepancy of binning bacteria and archaea, but not viruses, is a convention that likely hinders advancement in the field of viral metagenomics. Development of virus binning tools, such as vRhyme [124], will fuel this advancement.

Table 1. Recommendations for the questions, biases, and pitfalls posed in each section.

Sweeping contamination under the rug: balancing recovery and false discovery *All software tools that predict viruses from metagenomes can make mistakes*
1. Using multiple virus prediction tools and combining results can strengthen predictions by mitigating the biases and pitfall of each individual tool 2. In published work, report all parameters and thresholds used for predicting viruses, including methods of manual curation 3. Selecting low thresholds when running software or retaining low probability predictions will often generate "more data" at the expense of that data being low quality (i.e., contaminated) 4. Read the tool's publication (if available) in addition to the software documentation to best understand the tool's utility, pitfalls, and performance benchmarks
Of reference and reality *The reliance of most software tools on reference databases is a source of bias*
1. Consider homology search to additional curated databases in addition to NCBI databases when reporting novel sequences or gene features
The reference-free fallacy: no such thing as a reference-free virus prediction *No current tool for predicting virus sequences is reference-free*
1. Repeated training tools on NCBI databases has led to overlap in training and testing datasets across tools, making benchmarks increasingly difficult to perform without bias. Including non-NCBI databases in training, testing, and curating databases can reduce bias 2. Avoid falsely assuming database-independent machine learning models, whether trained on protein annotations or nucleotide features, overcome the necessity for reference-based searches
Linear genomes can be complete: where did all the linear genomes go? *Emphasis is placed on circular genomes as complete, excluding linear genomes*
1. Although complete, linear genomes may be identified as high quality or near complete, the lack of circularization signatures underemphasizes these genomes in databases or analyses 2. A metagenomics-scale approach to identify complete viral genomes without terminal repeats may reduce the bias towards circular genomes. Until such a tool is available, it is necessary to keep in mind the possibility of underrepresenting linear genomes
Metagenomes are puzzles: an unfinished puzzle is still just pieces *Not all metagenomic viral scaffolds represent the whole genome*
1. The inclusion of binning in virus analysis pipelines and constructing viral metagenome-assembled genomes (vMAGs) will likely better represent true composition of viruses and viral diversity

Conclusions

Virus genomics, specifically metagenomics, allows for the circumvention of conventional cultivation approaches to study viruses, their impacts on microbial communities, biogeochemistry, applications for biotechnology, human medicine, and more. After sequencing a sample, it has become just a few keystrokes and a click of a button to obtain a list of the viruses present. The outcome is that our knowledge of viral genomic diversity has increased at a near exponential rate over the last few years, opening new and exciting opportunities. However, this has been at the expense of biasing conclusions due to tools, methodologies, and conventions that lag data acquisition.

We are led to several overarching questions. Are virus predictions capturing the true nature of a community of viruses? Are heavily reference-guided predictions making it easy to miss any undiscovered novelty without studious inspection? Are conventions in identifying high-quality and complete viral genomes ignoring entire viral groups with unique genome architecture? Is the field as a whole moving too fast to fully consider the scope of the genomes presented?

There is no single set of answers to address all these questions easily. Rather, recognizing the limitations of the available methods will help to best work towards an optimized, efficient, and accurate approach to handle the rapid, near-constant flow of sequencing information. The goal is a fair, holistic representation of the global virosphere to best understand how viruses influence all life.

Acknowledgements

We thank Evelien Adriaenssens and Jelle Matthijnssens for the invitation to contribute to this special series. We thank members of the Anantharaman laboratory at the University of Wisconsin-

Madison for helpful feedback and discussions. K.K. was supported by a Wisconsin Distinguished Graduate Fellowship Award from the University of Wisconsin-Madison, and a William H. Peterson Fellowship Award from the Department of Bacteriology, University of Wisconsin-Madison. This research was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM143024. Figures were created with Biorender.com.

References

1. Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nature Reviews Microbiology* 1–14 (2020) doi:10.1038/s41579-019-0311-5.
2. Rohwer, F. Global Phage Diversity. *Cell* **113**, 141 (2003).
3. Comeau, A. M. *et al.* Exploring the prokaryotic virosphere. *Research in Microbiology* **159**, 306–313 (2008).
4. Koonin, E. V., Dolja, V. V., Krupovic, M. & Kuhn, J. H. Viruses Defined by the Position of the Virosphere within the Replicator Space. *Microbiology and Molecular Biology Reviews* **85**, e00193-20.
5. Wilhelm, S. W. & Suttle, C. A. Viruses and Nutrient Cycles in the Sea. *BioScience* **49**, 8 (1999).
6. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).
7. Berliner, A. J., Mochizuki, T. & Stedman, K. M. Astrovirology: Viruses at Large in the Universe. *Astrobiology* **18**, 207–223 (2018).
8. Suttle, C. A. Viruses: a vast reservoir of genetic diversity and driver of global processes. *Retrovirology* **6**, I7 (2009).
9. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
10. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
11. Mangalea, M. R. *et al.* Individuals at risk for rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential. *Cell Host & Microbe* **29**, 726-739.e5 (2021).
12. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host & Microbe* **26**, 527-541.e5 (2019).
13. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe* **26**, 764-778.e5 (2019).
14. Mumford, R. A., Macarthur, R. & Boonham, N. The role and challenges of new diagnostic technology in plant biosecurity. *Food Sec.* **8**, 103–109 (2016).
15. Rubio, L., Galipienso, L. & Ferriol, I. Detection of Plant Viruses and Disease Management: Relevance of Genetic Diversity and Evolution. *Frontiers in Plant Science* **11**, (2020).
16. Nasheri, N., Vester, A. & Petronella, N. Foodborne viral outbreaks associated with frozen produce. *Epidemiol Infect* **147**, e291 (2019).
17. Pujato, S. a., Quiberoni, A. & Mercanti, D. j. Bacteriophages on dairy foods. *Journal of Applied Microbiology* **126**, 14–30 (2019).
18. Wellenberg, G. J., van der Poel, W. H. M. & Van Oirschot, J. T. Viral infections and bovine mastitis: a review. *Veterinary Microbiology* **88**, 27–45 (2002).
19. Mi, S. *et al.* Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**, 785–789 (2000).
20. Blond, J.-L. *et al.* An Envelope Glycoprotein of the Human Endogenous Retrovirus HERV-W Is Expressed in the Human Placenta and Fuses Cells Expressing the Type D Mammalian Retrovirus Receptor. *J Virol* **74**, 3321–3329 (2000).
21. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proceedings of the National Academy of Sciences* **110**, 10771–10776 (2013).

22. Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *J Mol Evol* **60**, 174–182 (2005).
23. Pourcel, C., Salvignol, G. & Vergnaud, G. Y. 2005. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663.
24. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).
25. Milanesi, G. *et al.* BK virus-plasmid expression vector that persists episomally in human cells and shuttles into Escherichia coli. *Mol Cell Biol* **4**, 1551–1560 (1984).
26. Suzuki, M., Kasai, K. & Saeki, Y. Plasmid DNA Sequences Present in Conventional Herpes Simplex Virus Amplicon Vectors Cause Rapid Transgene Silencing by Forming Inactive Chromatin. *Journal of Virology* **80**, 3293–3300 (2006).
27. Hitt, M. M. & Graham, F. L. Adenovirus vectors for human gene therapy. *Adv Virus Res* **55**, 479–505 (2000).
28. Danthinne, X. & Imperiale, M. J. Production of first generation adenovirus vectors: a review. *Gene Ther* **7**, 1707–1714 (2000).
29. Fujimoto, K. *et al.* Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts. *Cell Host & Microbe* **28**, 380-389.e9 (2020).
30. Chatterjee, A. *et al.* Parallel Genomics Uncover Novel Enterococcal-Bacteriophage Interactions. *mBio* **11**, (2020).
31. Kortright, K. E., Chan, B. K., Koff, J. L. & Turner, P. E. Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host & Microbe* **25**, 219–232 (2019).
32. Sharma, R. S., Karmakar, S., Kumar, P. & Mishra, V. Application of filamentous phages in environment: A tectonic shift in the science and practice of ecorestoration. *Ecology and Evolution* **9**, 2263–2304 (2019).
33. Harada, L. K. *et al.* Biotechnological applications of bacteriophages: State of the art. *Microbiological Research* **212–213**, 38–58 (2018).
34. Schroven, K., Aertsen, A. & Lavigne, R. Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation. *FEMS Microbiology Reviews* **45**, (2021).
35. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
36. Hurwitz, B. L. & U'Ren, J. M. Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology* **31**, 161–168 (2016).
37. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801–812 (2007).
38. Hurwitz, B. L., Hallam, S. J. & Sullivan, M. B. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biology* **14**, R123 (2013).
39. Eggers, C. H. *et al.* Phage-mediated horizontal gene transfer of both prophage and heterologous DNA by ϕBB-1, a bacteriophage of Borrelia burgdorferi. *Pathog Dis* **74**, (2016).
40. Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* **17**, 930 (2016).

41. Jiang, S. C. & Paul, J. H. Gene Transfer by Transduction in the Marine Environment. *APPL. ENVIRON. MICROBIOL.* **64**, 8 (1998).
42. Touchon, M., Moura de Sousa, J. A. & Rocha, E. P. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Current Opinion in Microbiology* **38**, 66–73 (2017).
43. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA Bacteriophage Diversity. *PLOS Biology* **14**, e1002409 (2016).
44. Breitbart, M., Thompson, L., Suttle, C. & Sullivan, M. Exploring the Vast Diversity of Marine Viruses. *Oceanography* **20**, 135–139 (2007).
45. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral Mutation Rates. *J Virol* **84**, 9733–9748 (2010).
46. Peck, K. M. & Lauring, A. S. Complexities of Viral Mutation Rates. *J Virol* **92**, e01031-17 (2018).
47. Westra, E. R., Sünderhauf, D., Landsberger, M. & Buckling, A. Mechanisms and consequences of diversity-generating immune strategies. *Nat Rev Immunol* **17**, 719–728 (2017).
48. Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548 (1999).
49. Stern, A. & Sorek, R. The phage-host arms race: Shaping the evolution of microbes. *BioEssays* **33**, 43–51 (2011).
50. Tal, N. & Sorek, R. SnapShot: Bacterial immunity. *Cell* **185**, 578-578.e1 (2022).
51. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nature Reviews Microbiology* **18**, 113–119 (2020).
52. Tree, M. O. *et al.* Insect-specific flavivirus infection is restricted by innate immunity in the vertebrate host. *Virology* **497**, 81–91 (2016).
53. Mahmoudabadi, G., Milo, R. & Phillips, R. Energetic cost of building a virus. *PNAS* **114**, E4324–E4333 (2017).
54. Edwards, K. F., Steward, G. F. & Schvarcz, C. R. Making sense of virus size and the tradeoffs shaping viral fitness. *Ecology Letters* **n/a**,.
55. Bae, T., Baba, T., Hiramatsu, K. & Schneewind, O. Prophages of Staphylococcus aureus Newman and their contribution to virulence. *Molecular Microbiology* **62**, 1035–1047 (2006).
56. Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910–1914 (1996).
57. Heldal, M. & Bratbak, G. Production and decay of viruses in aquatic environments. *Mar. Ecol. Prog. Ser.* **72**, 205–212 (1991).
58. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography* **45**, 1320–1328 (2000).
59. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754–766 (2018).
60. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc Natl Acad Sci USA* **118**, e2023202118 (2021).

61. Luo, E., Eppley, J. M., Romano, A. E., Mende, D. R. & DeLong, E. F. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *ISME J* **14**, 1304–1315 (2020).
62. Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J. & Temperton, B. Host-hijacking and planktonic piracy: how phages command the microbial high seas. *Virol J* **16**, (2019).
63. Jover, L. F., Effler, T. C., Buchan, A., Wilhelm, S. W. & Weitz, J. S. The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nature Reviews Microbiology* **12**, 519–528 (2014).
64. Birch, E. W., Ruggero, N. A. & Covert, M. W. Determining Host Metabolic Limitations on Viral Replication via Integrated Modeling and Experimental Perturbation. *PLOS Computational Biology* **8**, e1002746 (2012).
65. Rosenwasser, S., Ziv, C., Creveld, S. G. van & Vardi, A. Virocell Metabolism: Metabolic Innovations During Host-Virus Interactions in the Ocean. *Trends Microbiol* **24**, 821–832 (2016).
66. Stent, G. S. & Maaløe, O. Radioactive phosphorus tracer studies on the reproduction of T4 bacteriophage: II. Kinetics of phosphorus assimilation. *Biochimica et Biophysica Acta* **10**, 55–69 (1953).
67. Cohen, S. S. The synthesis of bacterial viruses; the origin of the phosphorus found in the desoxyribonucleic acids of the T2 and T4 bacteriophages. *J Biol Chem* **174**, 295–303 (1948).
68. Kozloff, L. M., Knowlton, K., Putnam, F. W. & Evans, E. A. Biochemical Studies of Virus Reproduction V. the Origin of Bacteriophage Nitrogen. *J. Biol. Chem.* **188**, 101–116 (1951).
69. Waldbauer, J. R. *et al.* Nitrogen sourcing during viral infection of marine cyanobacteria. *PNAS* **116**, 15590–15595 (2019).
70. Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *ISME J* 1–15 (2020) doi:10.1038/s41396-019-0580-z.
71. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
72. Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nature Communications* **10**, 752 (2019).
73. Mann, N. H., Cook, A., Millard, A., Bailey, S. & Clokie, M. Bacterial photosynthesis genes in a virus. *Nature* **424**, 741 (2003).
74. Bragg, J. G. & Chisholm, S. W. Modeling the Fitness Consequences of a Cyanophage-Encoded Photosynthesis Gene. *PLOS ONE* **3**, e3550 (2008).
75. Thompson, L. R. *et al.* Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *PNAS* **108**, E757–E764 (2011).
76. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–89 (2005).
77. Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B. & Konstantinidis, K. T. Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environmental Microbiology Reports* **11**, 672–689 (2019).
78. Sullivan, M. B. *et al.* Prevalence and Evolution of Core Photosystem II Genes in Marine Cyanobacterial Viruses and Their Hosts. *PLoS Biology* **4**, e234 (2006).

79. Ahlgren, N. A., Fuchsman, C. A., Rocap, G. & Fuhrman, J. A. Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *The ISME Journal* **13**, 618–631 (2019).
80. Gazitúa, M. C. *et al.* Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *ISME J* **15**, 981–998 (2021).
81. Chen, L.-X. *et al.* Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat Microbiol* **5**, 1504–1515 (2020).
82. Anantharaman, K. *et al.* Sulfur Oxidation Genes in Diverse Deep-Sea Viruses. *Science* **344**, 757–760 (2014).
83. Maina, A. N. *et al.* Novel PhoH-encoding vibriophages with lytic activity against environmental Vibrio strains. *Arch Microbiol* **203**, 5321–5331 (2021).
84. Trubl, G. *et al.* Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* **3**, e00076-18 (2018).
85. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* **3**, 870 (2018).
86. Olsen, G. J., Lane, D. J., Giovannoni, S. J., Pace, N. R. & Stahl, D. A. Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* **40**, 337–365 (1986).
87. Schmidt, T. M., DeLong, E. F. & Pace, N. R. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* **173**, 4371–4378 (1991).
88. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics* **CHAPTER**, Unit-11.5 (2010).
89. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**, 455–477 (2012).
90. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
91. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
92. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, (2019).
93. Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**, 26 (2014).
94. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144–1146 (2014).
95. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* 1–6 (2021) doi:10.1038/s41587-020-00777-4.
96. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
97. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).
98. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
99. Kelley, D. R., Liu, B., Delcher, A. L., Pop, M. & Salzberg, S. L. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res* **40**, e9 (2012).

100. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res* **15**, 387–396 (2008).
101. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
102. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
103. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
104. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
105. Bushnell, B., Rood, J. & Singer, E. BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE* **12**, e0185056 (2017).
106. Santos-Medellin, C. *et al.* Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J* **15**, 1956–1970 (2021).
107. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of cells. *Biology Direct* **1**, 29 (2006).
108. Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A Roadmap for Genome-Based Phage Taxonomy. *Viruses* **13**, 506 (2021).
109. Hatfull, G. F. *et al.* Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* **397**, 119–143 (2010).
110. Philippe, N. *et al.* Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**, 281–286 (2013).
111. Zhan, Y. & Chen, F. The smallest ssDNA phage infecting a marine bacterium. *Environmental Microbiology* **21**, 1916–1928 (2019).
112. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *The ISME Journal* **11**, 1511–1520 (2017).
113. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
114. Auslander, N., Gussow, A. B., Benler, S., Wolf, Y. I. & Koonin, E. V. Seeker: alignment-free identification of bacteriophage genomes by deep learning. *Nucleic Acids Research* **48**, e121 (2020).
115. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, (2015).
116. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
117. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
118. Arisdakessian, C. G., Nigro, O. D., Steward, G. F., Poisson, G. & Belcaid, M. CoCoNet: an efficient deep learning tool for viral metagenome binning. *Bioinformatics* **37**, 2803–2810 (2021).
119. Johansen, J. *et al.* Genome binning of viral entities from bulk metagenomics data. *Nat Commun* **13**, 965 (2022).

120. Schulz, F. *et al.* Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. *mSystems* **5**, e00048-20 (2020).
121. Kieft, K. *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Reports* **36**, (2021).
122. Kieft, K. *et al.* Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat Commun* **12**, 3503 (2021).
123. Kieft, K. & Anantharaman, K. Deciphering active prophages from metagenomes. *bioRxiv* 2021.01.29.428894 (2021) doi:10.1101/2021.01.29.428894.
124. Kieft, K., Adams, A., Salamzade, R., Kalan, L. & Anantharaman, K. *vRhyme enables binning of viral genomes from metagenomes*. 2021.12.16.473018 https://www.biorxiv.org/content/10.1101/2021.12.16.473018v1 (2021) doi:10.1101/2021.12.16.473018.
125. Kieft, K. & Anantharaman, K. Virus genomics: what is being overlooked? *Current Opinion in Virology* **53**, 101200 (2022).
126. Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nature Geoscience* **4**, 698–702 (2011).
127. Fike, D. A., Bradley, A. S. & Rose, C. V. Rethinking the Ancient Sulfur Cycle. *Annual Review of Earth and Planetary Sciences* **43**, 593–622 (2015).
128. Andreae, M. O. Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. *Marine Chemistry* **30**, 1–29 (1990).
129. Voordouw, G. *et al.* Characterization of 16S rRNA genes from oil field microbial communities indicates the presence of a variety of sulfate-reducing, fermentative, and sulfide-oxidizing bacteria. *Appl Environ Microbiol* **62**, 1623–1629 (1996).
130. Ma, H. *et al.* The influence of hydrogen sulfide on corrosion of iron under different conditions. *Corrosion Science* **42**, 1669–1683 (2000).
131. Guo, F.-F., Yu, T.-C., Hong, J. & Fang, J.-Y. Emerging Roles of Hydrogen Sulfide in Inflammatory and Neoplastic Colonic Diseases. *Front Physiol* **7**, (2016).
132. Anantharaman, K. *et al.* Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *The ISME Journal* **12**, 1715 (2018).
133. Carbonero, F., Benefiel, A. C., Alizadeh-Ghamsari, A. H. & Gaskins, H. R. Microbial pathways in colonic sulfur metabolism and links with health and disease. *Front Physiol* **3**, (2012).
134. Morra, M. J. & Dick, W. A. Mechanisms of H2S Production from Cysteine and Cystine by Microorganisms Isolated from Soil by Selective Enrichment. *Appl Environ Microbiol* **57**, 1413–1417 (1991).
135. Xia, Y. *et al.* Sulfide production and oxidation by heterotrophic bacteria under aerobic conditions. *The ISME Journal* **11**, 2754–2766 (2017).
136. Gobler, C. J., Hutchins, D. A., Fisher, N. S., Cosper, E. M. & Saňudo-Wilhelmy, S. A. Release and bioavailability of C, N, P Se, and Fe following viral lysis of a marine chrysophyte. *Limnology and Oceanography* **42**, 1492–1504 (1997).
137. Jiao, N. *et al.* Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology* **8**, 593–599 (2010).
138. Manojlović, L. M. Photometry-based estimation of the total number of stars in the Universe. *Applied Optics* **54**, 6589 (2015).

139. Hurwitz, B. L., Brum, J. R. & Sullivan, M. B. Depth-stratified functional and taxonomic niche specialization in the 'core' and 'flexible' Pacific Ocean Virome. *ISME J* **9**, 472–484 (2015).
140. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife Sciences* **3**, e03125 (2014).
141. Paez-Espino, D. *et al.* IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
142. Linz, A. M. *et al.* Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ* **6**, e6075 (2018).
143. Rahlff, J. *et al.* Genome-informed microscopy reveals infections of uncultivated carbon-fixing archaea by lytic viruses in Earth's crust. *bioRxiv* 2020.07.22.215848 (2020) doi:10.1101/2020.07.22.215848.
144. Schulz, F. *et al.* Advantages and Limits of Metagenomic Assembly and Binning of a Giant Virus. *mSystems* **5**, (2020).
145. Fitzgerald, J. W. Sulfate ester formation and hydrolysis: a potentially important yet often ignored aspect of the sulfur cycle of aerobic soils. *Bacteriol Rev* **40**, 698–721 (1976).
146. Chiku, T. *et al.* H2S Biogenesis by Human Cystathionine γ-Lyase Leads to the Novel Sulfur Metabolites Lanthionine and Homolanthionine and Is Responsive to the Grade of Hyperhomocysteinemia. *J Biol Chem* **284**, 11601–11612 (2009).
147. Norman, J. M. *et al.* Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell* **160**, 447–460 (2015).
148. Henriques, A. C. & De Marco, P. Methanesulfonate (MSA) Catabolic Genes from Marine and Estuarine Bacteria. *PLOS ONE* **10**, e0125735 (2015).
149. Poyraz, Ö. *et al.* Crystal Structures of the Kinase Domain of the Sulfate-Activating Complex in Mycobacterium tuberculosis. *PLoS One* **10**, (2015).
150. Ishikawa, K., Mino, K. & Nakamura, T. New function and application of the cysteine synthase from archaea. *Biochemical Engineering Journal* **48**, 315–322 (2010).
151. Komoto, J., Yamada, T., Takata, Y., Markham, G. D. & Takusagawa, F. Crystal Structure of the S-Adenosylmethionine Synthetase Ternary Complex: A Novel Catalytic Mechanism of S-Adenosylmethionine Synthesis from ATP and Met,. *Biochemistry* **43**, 1821–1831 (2004).
152. Chartron, J. *et al.* Substrate Recognition, Protein Dynamics, and Iron-Sulfur Cluster in Pseudomonas aeruginosa Adenosine 5′-Phosphosulfate Reductase. *J Mol Biol* **364**, 152–169 (2006).
153. Knauer, S. H., Hartl-Spiegelhauer, O., Schwarzinger, S., Hänzelmann, P. & Dobbek, H. The Fe(II)/α-ketoglutarate-dependent taurine dioxygenases from Pseudomonas putida and Escherichia coli are tetramers. *The FEBS Journal* **279**, 816–831 (2012).
154. Sernova, N. V. & Gelfand, M. S. Identification of replication origins in prokaryotic genomes. *Brief Bioinform* **9**, 376–391 (2008).
155. Olm, M. R. *et al.* Identical bacterial populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in situ growth rates. *Genome Res.* (2017) doi:10.1101/gr.213256.116.
156. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nature Biotechnology* **34**, 1256–1263 (2016).

157. Liu, X., Jiang, H., Gu, Z. & Roberts, J. W. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proc Natl Acad Sci U S A* **110**, 11928–11933 (2013).
158. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 1 (2019) doi:10.1038/s41586-019-0965-1.
159. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* 1 (2019) doi:10.1038/s41586-019-1058-x.
160. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).
161. Li, Y., Hugenholtz, J., Sybesma, W., Abee, T. & Molenaar, D. Using Lactococcus lactis for glutathione overproduction. *Appl. Microbiol. Biotechnol.* **67**, 83–90 (2005).
162. Park, S. & Imlay, J. A. High Levels of Intracellular Cysteine Promote Oxidative DNA Damage by Driving the Fenton Reaction. *J. Bacteriol.* **185**, 1942–1950 (2003).
163. Fenner, L., Roux, V., Ananian, P. & Raoult, D. Alistipes finegoldii in Blood Cultures from Colon Cancer Patients. *Emerg Infect Dis* **13**, 1260–1262 (2007).
164. Hugon, P. *et al.* Non contiguous-finished genome sequence and description of Alistipes obesi sp. nov. *Stand Genomic Sci* **7**, 427–439 (2013).
165. Patrascu, O. *et al.* A fibrolytic potential in the human ileum mucosal microbiota revealed by functional metagenomic. *Sci Rep* **7**, (2017).
166. Schirmer, M. *et al.* Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nature Microbiology* **3**, 337–346 (2018).
167. Bakir, M. A., Kitahara, M., Sakamoto, M., Matsumoto, M. & Benno, Y. Bacteroides intestinalis sp. nov., isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology* **56**, 151–154 (2006).
168. Costea, P. I. *et al.* Subspecies in the global human gut microbiome. *Mol Syst Biol* **13**, (2017).
169. Curtis, M. M. *et al.* The Gut Commensal Bacteroides thetaiotaomicron Exacerbates Enteric Infection through Modification of the Metabolic Landscape. *Cell Host & Microbe* **16**, 759–769 (2014).
170. Jiang, W. *et al.* Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. *Scientific Reports* **5**, 8096 (2015).
171. Kuang, Y.-S. *et al.* Connections between the human gut microbiome and gestational diabetes mellitus. *Gigascience* **6**, (2017).
172. Martín, R. *et al.* Functional Characterization of Novel Faecalibacterium prausnitzii Strains Isolated from Healthy Volunteers: A Step Forward in the Use of F. prausnitzii as a Next-Generation Probiotic. *Front Microbiol* **8**, (2017).
173. Pfleiderer, A. *et al.* Non-contiguous finished genome sequence and description of Alistipes ihumii sp. nov. *Standards in Genomic Sciences* **9**, 1221 (2014).
174. Qin, J. *et al.* A human gut microbial gene catalog established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
175. Veiga, P. *et al.* Changes of the human gut microbiome induced by a fermented milk product. *Sci Rep* **4**, (2014).

176. Eeckhaut, V. *et al.* Butyricicoccus pullicaecorum in inflammatory bowel disease. *Gut* **62**, 1745–1752 (2013).
177. Larsen, J. M. The immune response to Prevotella bacteria in chronic inflammatory disease. *Immunology* **151**, 363–374 (2017).
178. Lucke, K. Prevalence of Bacteroides and Prevotella spp. in ulcerative colitis. *Journal of Medical Microbiology* **55**, 617–624 (2006).
179. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nature Microbiology* **4**, 693–700 (2019).
180. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications* **5**, 4498 (2014).
181. Villion, M. *et al.* P087, a lactococcal phage with a morphogenesis module similar to an Enterococcus faecalis prophage. *Virology* **388**, 49–56 (2009).
182. Moran, M. A. & Durham, B. P. Sulfur metabolites in the pelagic ocean. *Nature Reviews Microbiology* **17**, 665–678 (2019).
183. Gyaneshwar, P. *et al.* Sulfur and Nitrogen Limitation in Escherichia coli K-12: Specific Homeostatic Responses. *Journal of Bacteriology* **187**, 1074–1090 (2005).
184. Nambi, S. *et al.* The oxidative stress network of Mycobacterium tuberculosis reveals coordination between radical detoxification systems. *Cell Host Microbe* **17**, 829–837 (2015).
185. Pal, V. K., Bandyopadhyay, P. & Singh, A. Hydrogen Sulfide in Physiology and Pathogenesis of Bacteria and Viruses. *IUBMB Life* **70**, 393–410 (2018).
186. Peng, H. *et al.* Sulfide Homeostasis and Nitroxyl Intersect via Formation of Reactive Sulfur Species in Staphylococcus aureus. *mSphere* **2**, (2017).
187. Tam, W. *et al.* Tail tip proteins related to bacteriophage λ gpL coordinate an iron-sulfur cluster. *J. Mol. Biol.* **425**, 2450–2462 (2013).
188. Damon, J. R., Pincus, D. & Ploegh, H. L. tRNA thiolation links translation to stress responses in Saccharomyces cerevisiae. *Mol Biol Cell* **26**, 270–282 (2015).
189. Hsu, W. T., Foft, J. W. & Weiss, S. B. Effect of bacteriophage infection on the sulfur-labeling of sRNA. *Proc Natl Acad Sci U S A* **58**, 2028–2035 (1967).
190. Lira, N. P. V. de *et al.* BigR is a sulfide sensor that regulates a sulfur transferase/dioxygenase required for aerobic respiration of plant bacteria under sulfide stress. *Scientific Reports* **8**, 3508 (2018).
191. Shimizu, T. *et al.* Sulfide-responsive transcriptional repressor SqrR functions as a master regulator of sulfide-dependent photosynthesis. *Proceedings of the National Academy of Sciences* **114**, 2355–2360 (2017).
192. Yang, Y. *et al.* DNA Backbone Sulfur-Modification Expands Microbial Growth Range under Multiple Stresses by its anti-oxidation function. *Scientific Reports* **7**, 3516 (2017).
193. Jessop, L., Bankhead, T., Wong, D. & Segall, A. M. The Amino Terminus of Bacteriophage Integrase Is Involved in Protein-Protein Interactions during Recombination. *J. BACTERIOL.* **182**, 11 (2000).
194. Kessler, D. Enzymatic activation of sulfur for incorporation into biomolecules in prokaryotes. *FEMS Microbiol Rev* **30**, 825–840 (2006).
195. Yeeles, J. T. P., Cammack, R. & Dillingham, M. S. An Iron-Sulfur Cluster Is Essential for the Binding of Broken DNA by AddAB-type Helicase-Nucleases. *J Biol Chem* **284**, 7746–7755 (2009).

196. Propst-Ricciuti, C. The Effect of Host-Cell Starvation on Virus-induced Lysis by MS2 Bacteriophage. *Journal of General Virology* **31**, 323–330 (1976).
197. Yin, J. *et al.* l-Cysteine metabolism and its nutritional implications. *Molecular Nutrition & Food Research* **60**, 134–146 (2016).
198. Habicht, K. S. & Canfield, D. E. Sulfur isotope fractionation during bacterial sulfate reduction in organic-rich sediments. *Geochimica et Cosmochimica Acta* **61**, 5351–5361 (1997).
199. Sim, M. S. *et al.* Role of APS reductase in biogeochemical sulfur isotope fractionation. *Nature Communications* **10**, 44 (2019).
200. Thode, H. G., Macnamara, J. & Fleming, W. H. Sulphur isotope fractionation in nature and geological and biological time scales. *Geochimica et Cosmochimica Acta* **3**, 235–243 (1953).
201. Kaplan, I. R. & Rittenberg, S. C. Microbiological Fractionation of Sulphur Isotopes. *Journal of General Microbiology* **34**, 195–212 (1964).
202. Chambers, L. A. & Trudinger, P. A. Microbiological fractionation of stable sulfur isotopes: A review and critique. *Geomicrobiology Journal* **1**, 249–293 (1979).
203. Durfee, T. *et al.* The Complete Genome Sequence of Escherichia coli DH10B: Insights into the Biology of a Laboratory Workhorse. *Journal of Bacteriology* **190**, 2597–2606 (2008).
204. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
205. Markowitz, V. M. *et al.* IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* **42**, D568–D573 (2014).
206. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
207. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).
208. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571-577 (2015).
209. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44**, 6614–6624 (2016).
210. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **44**, D67–D72 (2016).
211. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**, D200–D203 (2017).
212. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
213. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699–2699 (2018).
214. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 772–780 (2013).
215. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
216. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

217. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).
218. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
219. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* **11**, 2864–2868 (2017).
220. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
221. Rambaut, A. FigTree version 1.4.3. (2009).
222. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
223. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **40**, D742-753 (2012).
224. Linz, A. M., Aylward, F. O., Bertilsson, S. & McMahon, K. D. Time-series metatranscriptomes reveal conserved patterns between phototrophic and heterotrophic microbes in diverse freshwater systems. *Limnology and Oceanography* **65**, S101–S112 (2020).
225. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
226. Lillehaug, D. An improved plaque assay for poor plaque-producing temperate lactococcal bacteriophages. *Journal of Applied Microbiology* **83**, 85–90 (1997).
227. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: a genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
228. Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **9**, 90–95 (2007).
229. Ando, H., Lemire, S., Pires, D. P. & Lu, T. K. Engineering Modular Viral Scaffolds for Targeted Bacterial Population Editing. *cels* **1**, 187–196 (2015).
230. Jaschke, P. R., Lieberman, E. K., Rodriguez, J., Sierra, A. & Endy, D. A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology* **434**, 278–284 (2012).
231. Kuijpers, N. G. *et al.* A versatile, efficient strategy for assembly of multi-fragment expression vectors in Saccharomyces cerevisiae using 60 bp synthetic recombination sequences. *Microbial Cell Factories* **12**, 47 (2013).
232. Daniel Gietz, R. & Woods, R. A. Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. in *Methods in Enzymology* (eds. Guthrie, C. & Fink, G. R.) vol. 350 87–96 (Academic Press, 2002).
233. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408 (2001).
234. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45 (2011).
235. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev* **40**, 258–272 (2016).
236. Louca, S., Mazel, F., Doebeli, M. & Parfrey, L. W. A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biology* **17**, e3000106 (2019).
237. Russell, P. W. & Müller, U. R. Construction of bacteriophage luminal diameterX174 mutants with maximum genome sizes. *J Virol* **52**, 822–827 (1984).

238. Lindell, D. *et al.* Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013–11018 (2004).
239. Lindell, D. *et al.* Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* **449**, 83–86 (2007).
240. Cassman, N. *et al.* Oxygen minimum zones harbour novel viral communities with low diversity: Viral community characteristics of an oxygen minimum zone. *Environmental Microbiology* **14**, 3043–3065 (2012).
241. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat Commun* **8**, 15892 (2017).
242. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their Genomes. *Curr Opin Virol* **1**, 298–303 (2011).
243. Ikeuchi, Y., Shigi, N., Kato, J., Nishimura, A. & Suzuki, T. Mechanistic Insights into Sulfur Relay by Multiple Sulfur Mediators Involved in Thiouridine Biosynthesis at tRNA Wobble Positions. *Molecular Cell* **21**, 97–108 (2006).
244. Dammeyer, T., Bagby, S., Sullivan, M., Chisholm, S. & Frankenberg-Dinkel, N. Efficient phage-mediated pigment biosynthesis in oceanic cyanobacteria. *Curr Biol* **18**, (2008).
245. Ghosh, W. & Dam, B. Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiol Rev* **33**, 999–1043 (2009).
246. Marshall, K. T. & Morris, R. M. Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. *The ISME Journal* **7**, 452–455 (2013).
247. Grimm, F., Dobler, N. & Dahl, C. Regulation of dsr genes encoding proteins responsible for the oxidation of stored sulfur in Allochromatium vinosum. *Microbiology* **156**, 764–773 (2010).
248. Bradley, A. S., Leavitt, W. D. & Johnston, D. T. Revisiting the dissimilatory sulfate reduction pathway. *Geobiology* **9**, 446–457 (2011).
249. Hensen, D., Sperling, D., Trüper, H. G., Brune, D. C. & Dahl, C. Thiosulphate oxidation in the phototrophic sulphur bacterium Allochromatium vinosum. *Mol. Microbiol.* **62**, 794–810 (2006).
250. Friedrich, C. G. *et al.* Novel genes coding for lithotrophic sulfur oxidation of Paracoccus pantotrophus GB17. *J. Bacteriol.* **182**, 4677–4687 (2000).
251. Hatfull, G. F. Bacteriophage Genomics. *Curr Opin Microbiol* **11**, 447–453 (2008).
252. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
253. Anantharaman, K., Breier, J. A. & Dick, G. J. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *The ISME Journal* **10**, 225–239 (2016).
254. Proctor, L. M., Okubo, A. & Fuhrman, J. A. Calibrating estimates of phage-induced mortality in marine bacteria: Ultrastructural studies of marine bacteriophage development from one-step growth experiments. *Microb Ecol* **25**, 161–182 (1993).
255. Hennes, K. P. & Simon, M. Significance of bacteriophages for controlling bacterioplankton growth in a mesotrophic lake. *Appl. Environ. Microbiol.* **61**, 333–340 (1995).
256. Anantharaman, K., Breier, J. A., Sheik, C. S. & Dick, G. J. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *PNAS* (2012) doi:10.1073/pnas.1215340110.

257. Zimmerman, A. E. *et al.* Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nature Reviews Microbiology* **18**, 21–34 (2020).
258. Breitbart, M. Marine Viruses: Truth or Dare. *Annual Review of Marine Science* **4**, 425–448 (2012).
259. Haveman, S. A. *et al.* Gene Expression Analysis of Energy Metabolism Mutants of Desulfovibrio vulgaris Hildenborough Indicates an Important Role for Alcohol Dehydrogenase. *Journal of Bacteriology* **185**, 4345–4353 (2003).
260. Puxty, R. J., Evans, D. J., Millard, A. D. & Scanlan, D. J. Energy limitation of cyanophage development: implications for marine carbon cycling. *The ISME Journal* **12**, 1273–1286 (2018).
261. Von Damm, K. L., Edmond, J. M., Measures, C. I. & Grant, B. Chemistry of submarine hydrothermal solutions at Guaymas Basin, Gulf of California. *Geochimica et Cosmochimica Acta* **49**, 2221–2237 (1985).
262. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686 (2019).
263. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).
264. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944 (2018).
265. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59–60 (2015).
266. Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* 1 (2019) doi:10.1038/s41587-019-0100-8.
267. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).
268. Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007).
269. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, (2020).
270. Michael Waskom *et al. mwaskom/seaborn: v0.8.1 (September 2017)*. (Zenodo, 2017). doi:10.5281/zenodo.883859.
271. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).
272. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
273. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
274. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
275. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle (2011).
276. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

277. Breitbart, M. & Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends in Microbiology* **13**, 278–284 (2005).
278. Wommack, K. E. & Colwell, R. R. Virioplankton: Viruses in Aquatic Ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
279. Danovaro, R. & Serresi, M. Viral Density and Virus-to-Bacterium Ratio in Deep-Sea Sediments of the Eastern Mediterranean. *Appl. Environ. Microbiol.* **66**, 1857–1861 (2000).
280. Brussaard, C. P. D. *et al.* Global-scale processes with a nanoscale drive: the role of marine viruses. *The ISME Journal* **2**, 575–578 (2008).
281. Barr, J. J. Missing a Phage: Unraveling Tripartite Symbioses within the Human Gut. *mSystems* **4**, e00105-19 (2019).
282. Kim, B. *et al.* Phage-Derived Antibacterials: Harnessing the Simplicity, Plasticity, and Diversity of Phages. *Viruses* **11**, (2019).
283. Peng, S.-Y. *et al.* Highly potent antimicrobial modified peptides derived from the Acinetobacter baumannii phage endolysin LysAB2. *Sci Rep* **7**, 1–12 (2017).
284. Holt, A. *et al.* Phage-encoded cationic antimicrobial peptide used for outer membrane disruption in lysis. *bioRxiv* 515445 (2019) doi:10.1101/515445.
285. Labonté, J. M. *et al.* Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J* **9**, 2386–2399 (2015).
286. Trubl, G. *et al.* Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ* **4**, e1999 (2016).
287. Wommack, K. E. *et al.* VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6**, 427 (2012).
288. Roux, S. *et al.* Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074–3075 (2011).
289. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432 (2019).
290. Fang, Z. *et al.* PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8**, (2019).
291. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* **45**, 39–53 (2017).
292. Ponsero, A. J. & Hurwitz, B. L. The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes. *Front. Microbiol.* **10**, (2019).
293. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).
294. Zheng, T. *et al.* Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* **7**, 42 (2019).
295. Arndt, D. *et al.* PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* **44**, W16–W21 (2016).
296. Song, W. *et al.* Prophage Hunter: an integrative hunting tool for active prophages. *Nucleic Acids Res* **47**, W74–W80 (2019).
297. Merchant, N. *et al.* The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLOS Biology* **14**, e1002342 (2016).
298. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

299. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
300. He, Q. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1–11 (2017).
301. Delcher, A. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30**, 2478–2483 (2002).
302. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
303. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).
304. Ijaz, U. Z. *et al.* The distinct features of microbial 'dysbiosis' of Crohn's disease do not occur to the same extent in their unaffected, genetically-linked kindred. *PLoS ONE* **12**, e0172605 (2017).
305. Kristensen, D. M. *et al.* Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. *Journal of Bacteriology* **195**, 941–950 (2013).
306. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**, D491–D498 (2017).
307. Hendricks, S. P. & Mathews, C. K. Regulation of T4 Phage Aerobic Ribonucleotide Reductase: SIMULTANEOUS ASSAY OF THE FOUR ACTIVITIES. *J. Biol. Chem.* **272**, 2861–2865 (1997).
308. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **37**, 29–37 (2019).
309. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725–731 (2017).
310. Tucker, K. P., Parsons, R., Symonds, E. M. & Breitbart, M. Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *ISME J* **5**, 822–830 (2011).
311. Payet, J. P. & Suttle, C. A. To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnology and Oceanography* **58**, 465–474 (2013).
312. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* (2019) doi:10.1016/j.cell.2019.03.040.
313. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology* **13**, R79 (2012).
314. Strauss, J. *et al.* Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. *Inflamm. Bowel Dis.* **17**, 1971–1978 (2011).
315. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr Opin Gastroenterol* **31**, 69–75 (2015).
316. Clemente, J. C., Ursell, L. K., Parfrey, L. W. & Knight, R. The Impact of the Gut Microbiota on Human Health: An Integrative View. *Cell* **148**, 1258–1270 (2012).
317. Minot, S. S. & Willis, A. D. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. *Microbiome* **7**, 110 (2019).

318. Nishio, M., Okada, N., Miki, T., Haneda, T. & Danbara, H. Identification of the outer-membrane protein PagC required for the serum resistance phenotype in Salmonella enterica serovar Choleraesuis. *Microbiology (Reading, Engl.)* **151**, 863–873 (2005).
319. Ragunathan, P. T. & Vanderpool, C. K. Cryptic-Prophage-Encoded Small Protein DicB Protects Escherichia coli from Phage Infection by Inhibiting Inner Membrane Receptor Proteins. *Journal of Bacteriology* **201**, (2019).
320. Rasko, D. A. *et al.* The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates. *Journal of Bacteriology* **190**, 6881–6893 (2008).
321. Touchon, M., Bernheim, A. & Rocha, E. P. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J* **10**, 2744–2754 (2016).
322. Cochran, P., Kellogg, C. & Paul, J. Prophage induction of indigenous marine lysogenic bacteria by environmental pollutants. *Mar. Ecol. Prog. Ser.* **164**, 125–133 (1998).
323. Casjens, S. R. & Hendrix, R. W. Bacteriophage lambda: Early pioneer and still relevant. *Virology* **479–480**, 310–330 (2015).
324. Carrolo, M., Frias, M. J., Pinto, F. R., Melo-Cristino, J. & Ramirez, M. Prophage Spontaneous Activation Promotes DNA Release Enhancing Biofilm Formation in Streptococcus pneumoniae. *PLOS ONE* **5**, e15678 (2010).
325. Binnenkade, L., Teichmann, L. & Thormann, K. M. Iron Triggers λSo Prophage Induction and Release of Extracellular DNA in Shewanella oneidensis MR-1 Biofilms. *Appl Environ Microbiol* **80**, 5304–5316 (2014).
326. Feiner, R. *et al.* A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nature Reviews Microbiology* **13**, 641–650 (2015).
327. Wendling, C. C., Refardt, D. & Hall, A. R. Fitness benefits to bacteria of carrying prophages and prophage-encoded antibiotic-resistance genes peak in different environments. *Evolution* **n/a**,.
328. Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brüssow, H. Prophage Genomics. *Microbiol Mol Biol Rev* **67**, 238–276 (2003).
329. Banks, D. J., Lei, B. & Musser, J. M. Prophage induction and expression of prophage-encoded virulence factors in group A Streptococcus serotype M3 strain MGAS315. *Infect Immun* **71**, 7079–7086 (2003).
330. Dedrick, R. M. *et al.* Prophage-mediated defence against viral attack and viral counter-defence. *Nature Microbiology* **2**, 1–13 (2017).
331. Thingstad, T. & Lignell, R. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* **13**, 19–27 (1997).
332. Noguchi, Y. & Katayama, T. The Escherichia coli Cryptic Prophage Protein YfdR Binds to DnaA and Initiation of Chromosomal Replication Is Inhibited by Overexpression of the Gene Cluster yfdQ-yfdR-yfdS-yfdT. *Front Microbiol* **7**, (2016).
333. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat Commun* **1**, 1–9 (2010).
334. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nature Communications* **9**, 4781 (2018).
335. Jenkins, S. T., Beard, J. P. & Howe, T. G. B. Male-Specific Bacteriophage MS2 Propagation in Fluorophenylalanine-Resistant Escherichia coli K12. *Journal of Virology* **14**, 50–55 (1974).

336. Weed, L. L. & Cohen, S. S. The utilization of host pyrimidines in the synthesis of bacterial viruses. *J Biol Chem* **192**, 693–700 (1951).
337. Zborowsky, S. & Lindell, D. Resistance in marine cyanobacteria differs against specialist and generalist cyanophages. *PNAS* **116**, 16899–16908 (2019).
338. Waller, A. S. *et al.* Classification and quantification of bacteriophage taxa in human gut metagenomes. *The ISME Journal* **8**, 1391–1402 (2014).
339. Hertel, R. *et al.* Genome-Based Identification of Active Prophage Regions by Next Generation Sequencing in Bacillus licheniformis DSM13. *PLOS ONE* **10**, e0120759 (2015).
340. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
341. Skovgaard, O., Bak, M., Løbner-Olesen, A. & Tommerup, N. Genome-wide detection of chromosomal rearrangements, indels, and mutations in circular chromosomes by short read sequencing. *Genome Res.* **21**, 1388–1393 (2011).
342. Kleiner, M., Bushnell, B., Sanderson, K. E., Hooper, L. V. & Duerkop, B. A. Transductomics: sequencing-based detection and analysis of transduced DNA in pure cultures and microbial communities. *Microbiome* **8**, 158 (2020).
343. Gutiérrez, R. *et al.* Prophage-Driven Genomic Structural Changes Promote Bartonella Vertical Evolution. *Genome Biology and Evolution* **10**, 3089–3103 (2018).
344. Ho, C.-H., Stanton-Cook, M., Beatson, S. A., Bansal, N. & Turner, M. S. Stability of active prophages in industrial Lactococcus lactis strains in the presence of heat, acid, osmotic, oxidative and antibiotic stressors. *International Journal of Food Microbiology* **220**, 26–32 (2016).
345. Turkington, C. J. R., Abadi, N. N., Edwards, R. A. & Grasis, J. A. *hafeZ: Active prophage identification through read mapping*. http://biorxiv.org/lookup/doi/10.1101/2021.07.21.453177 (2021) doi:10.1101/2021.07.21.453177.
346. Kim, M.-S. & Bae, J.-W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J* **12**, 1127–1141 (2018).
347. Gasparrini, A. J. *et al.* Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nature Microbiology* **4**, 2285–2297 (2019).
348. Woodcroft, B. J. *et al.* Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**, 49–54 (2018).
349. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nature Communications* **6**, 1–13 (2015).
350. Wallace, J. L., Vong, L., McKnight, W., Dicay, M. & Martin, G. R. Endogenous and exogenous hydrogen sulfide promotes resolution of colitis in rats. *Gastroenterology* **137**, 569–578, 578.e1 (2009).
351. Kieft, K. *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep* **36**, 109471 (2021).
352. Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N. & Novick, R. P. The gene for toxic shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus. *Molecular Microbiology* **29**, 527–543 (1998).
353. Freeman, V. J. STUDIES ON THE VIRULENCE OF BACTERIOPHAGE-INFECTED STRAINS OF CORYNEBACTERIUM DIPHTHERIAE1. *J Bacteriol* **61**, 675–688 (1951).

354. Brüssow, H., Canchaya, C. & Hardt, W.-D. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).
355. Sakaguchi, Y. *et al.* The genome sequence of Clostridium botulinum type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. *Proc Natl Acad Sci U S A* **102**, 17472–17477 (2005).
356. Riedel, T. *et al.* A Clostridioides difficile bacteriophage genome encodes functional binary toxin-associated genes. *Journal of Biotechnology* **250**, 23–28 (2017).
357. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Academic Press, 2013).
358. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLOS ONE* **5**, e11147 (2010).
359. Drew, G. C., Stevens, E. J. & King, K. C. Microbial evolution and transitions along the parasite–mutualist continuum. *Nat Rev Microbiol* **19**, 623–638 (2021).
360. Roossinck, M. J. Move Over, Bacteria! Viruses Make Their Mark as Mutualistic Microbial Symbionts. *Journal of Virology* **89**, 6532–6535.
361. Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology* 1 (2019) doi:10.1038/s41564-019-0510-x.
362. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).
363. Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 24175 (2016).
364. Wang, Z., Wang, Z., Lu, Y. Y., Sun, F. & Zhu, S. SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* **35**, 4229–4238 (2019).
365. Mallawaarachchi, V., Wickramarachchi, A. & Lin, Y. GraphBin: refined binning of metagenomic contigs using assembly graphs. *Bioinformatics* **36**, 3307–3313 (2020).
366. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**, 836–843 (2018).
367. Graham, E. D., Heidelberg, J. F. & Tully, B. J. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* **5**, e3035 (2017).
368. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res* **28**, 569–580 (2018).
369. Siranosian, B. *et al.* Tetranucleotide usage highlights genomic heterogeneity among mycobacteriophages. *F1000Res* **4**, 36 (2015).
370. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578–585 (2021).
371. Li, M. *et al.* Genomic and transcriptomic evidence for scavenging of diverse organic compounds by widespread deep-sea archaea. *Nat Commun* **6**, 8933 (2015).
372. Tran, P. Q. *et al.* Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J* **15**, 1971–1986 (2021).

373. Okazaki, Y., Nishimura, Y., Yoshida, T., Ogata, H. & Nakano, S. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environmental Microbiology* **21**, 4740–4754 (2019).
374. Coutinho, F. H. *et al.* New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* **8**, 163 (2020).
375. Trubl, G. *et al.* Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7**, e7265 (2019).
376. Swaney, M. H., Sandstrom, S. & Kalan, L. R. *Cobamide sharing drives skin microbiome dynamics*. 2020.12.02.407395 https://www.biorxiv.org/content/10.1101/2020.12.02.407395v2 (2021) doi:10.1101/2020.12.02.407395.
377. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
378. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
379. Israeli, O. *et al.* Complete Genome Sequence of the First Camelpox Virus Case Diagnosed in Israel. *Microbiol Resour Announc* **8**, e00671-19 (2019).
380. Caro-Vegas, C. *et al.* Runaway Kaposi Sarcoma-associated Herpesvirus Replication correlates with systemic IL-10 Levels. *Virology* **539**, 18–25 (2020).
381. Kieft, K. & Anantharaman, K. Deciphering Active Prophages from Metagenomes. *mSystems* **0**, e00084-22.
382. Waskom, M. L. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).
383. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* **9**, 471–472 (2012).
384. Byrd, A. L., Belkaid, Y. & Segre, J. A. The human skin microbiome. *Nat Rev Microbiol* **16**, 143–155 (2018).
385. Chatterjee, A. *et al.* Parallel Genomics Uncover Novel Enterococcal-Bacteriophage Interactions. *mBio* **11**, e03120-19.
386. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).
387. Paez-Espino, D. *et al.* Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* **7**, 157 (2019).
388. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host & Microbe* **28**, 724-740.e8 (2020).
389. Saw, A. K. *et al.* Alignment-free method for DNA sequence clustering using Fuzzy integral similarity. *Scientific Reports* **9**, 3753 (2019).
390. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).
391. Deaton, J., Yu, F. B. & Quake, S. R. PhaMers identifies novel bacteriophage sequences from thermophilic hot springs. *bioRxiv* 169672 (2017) doi:10.1101/169672.
392. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

393. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Frontiers in Genetics* **9**, 304 (2018).
394. Aylward, F. O. & Moniruzzaman, M. ViralRecall—A Flexible Command-Line Tool for the Detection of Giant Virus Signatures in 'Omic Data. *Viruses* **13**, 150 (2021).
395. Ponsero, A. J. & Hurwitz, B. L. The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes. *Frontiers in Microbiology* **10**, 806 (2019).
396. Roux, S., Krupovic, M., Debroas, D., Forterre, P. & Enault, F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology* **3**, 130160.
397. Pratama, A. A. *et al.* Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. *PeerJ* **9**, e11447 (2021).
398. Ecale Zhou, C. L. *et al.* multiPhATE: bioinformatics pipeline for functional annotation of phage isolates. *Bioinformatics* **35**, 4402–4404 (2019).
399. Roux, S. *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Research* **49**, D764–D775 (2021).
400. Auslander, N., Gussow, A. B. & Koonin, E. V. Incorporating Machine Learning into Established Bioinformatics Frameworks. *International Journal of Molecular Sciences* **22**, 2903 (2021).
401. Zayed, A. A. *et al.* efam: an expanded, metaproteome-supported HMM profile database of viral protein families. *Bioinformatics* btab451 (2021) doi:10.1093/bioinformatics/btab451.
402. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* **48**, 8883–8900 (2020).
403. Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature; London* **554**, 118-122,122A-122T (2018).
404. Callanan, J. *et al.* Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances* (2020) doi:10.1126/sciadv.aay5981.
405. Casjens, S. R. & Gilcrease, E. B. Determining DNA Packaging Strategy by Analysis of the Termini of the Chromosomes in Tailed-Bacteriophage Virions. *Methods Mol Biol* **502**, 91–111 (2009).
406. Beaulaurier, J. *et al.* Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res.* **30**, 437–446 (2020).
407. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
408. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* 1–6 (2021) doi:10.1038/s41587-020-00777-4.
409. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
410. Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).
411. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun* **11**, 1710 (2020).