PHAGE: Therapy, Applications, and Research
Volume 2, Number 4, 2021
© Mary Ann Liebert, Inc.
DOI: 10.1089/phage.2021.0007

INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes

Ryan Cook, BSc,[1] Nathan Brown, PhD,[2] Tamsin Redgwell, PhD,[3] Branko Rihtman, PhD,[4] Megan Barnes, BSc,[2] Martha Clokie, PhD,[2] Dov J. Stekel, PhD,[5] Jon Hobman, PhD,[5] Michael A. Jones, PhD,[1] and Andrew Millard, PhD[2,i]

Abstract

Background: With advances in sequencing technology and decreasing costs, the number of phage genomes that have been sequenced has increased markedly in the past decade.

Materials and Methods: We developed an automated retrieval and analysis system for phage genomes (https://github.com/RyanCook94/inphared) to produce the INfrastructure for a PHAge REference Database (INPHARED) of phage genomes and associated metadata.

Results: As of January 2021, 14,244 complete phage genomes have been sequenced. The INPHARED data set is dominated by phages that infect a small number of bacterial genera, with 75% of phages isolated on only 30 bacterial genera. There is further bias, with significantly more lytic phage genomes (∼ 70%) than temperate (∼ 30%) within our database. Collectively, this results in ∼ 54% of temperate phage genomes originating from just three host genera. With much debate on the carriage of antibiotic resistance genes and their potential safety in phage therapy, we searched for putative antibiotic resistance genes. Frequency of antibiotic resistance gene carriage was found to be higher in temperate phages than in lytic phages and again varied with host.

Conclusions: Given the bias of currently sequenced phage genomes, we suggest to fully understand phage diversity, efforts should be made to isolate and sequence a larger number of phages, in particular temperate phages, from a greater diversity of hosts.

Keywords: phage genomes, antibiotic resistance genes, virulence genes, jumbo phages

Introduction

Bacteriophages (hereafter phages) are viruses that specifically infect bacteria and are thought to be the most abundant biological entities in the biosphere.[1] Phages may be obligately lytic (hereafter lytic) or temperate, whereby they have access to both the lytic and lysogenic cycle. Phages have many roles; in the oceans they are important in diverting the flow of carbon into dissolved and particulate organic matter through the lysis of their hosts,[1] or directly halting the fixation of CO_2 carried out by their cyanobacterial hosts.[2] In the human microbiome, it is becoming increasingly clear that phages play roles in the severity and symptoms of several diseases. Many recent studies have shown disease-specific alterations can be seen in the gut virome community in both gastrointestinal and systemic conditions, including irritable bowel disease,[3] AIDS,[4] malnutrition,[5] and diabetes.[6]

Phages alter the physiology of their bacterial hosts such as by causing increased virulence, a notable example being phage CTX that actually encodes the toxins within the genome of *Vibrio cholerae*, which cause cholera.[7] Furthermore, there are many cases where the expression of phage-encoded toxins cause otherwise harmless commensal bacteria to convert into a pathogen, including multidrug-resistant ST11 strains of *Pseudomonas aeruginosa*,[8,9] and the Shiga-toxin encoding *Escherichia coli.*[10] As well as increasing the virulence of host bacteria, phages can also utilize parts of their genomes known as auxiliary metabolic genes, homologues of

[1] School of Veterinary Medicine and Science, University of Nottingham, Loughborough, United Kingdom.
[2] Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom.
[3] COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark.
[4] School of Life Sciences, University of Warwick, Coventry, United Kingdom.
[5] School of Biosciences, University of Nottingham, Loughborough, United Kingdom.
[i] ORCID ID (https://orcid.org/0000-0002-3895-2854).

host metabolic genes, to modulate their host's metabolism that can again have profound impacts on bacterial physiology and disease.[11]

Our understanding of how phages alter host metabolism has increased as the number of phage genomes has been sequenced. The first phage genome in 1977,[12] and since then, the relative ease of high-throughput sequencing combined with the resurgence of interest in this topic, has led to a rapid increase in the number of sequenced phage genomes.[13,14] The relatively simple nature of phage genomes means that the vast majority of isolated phage genomes can be fully assembled using short-read next-generation sequencing approaches.[15] As temperate phages can integrate into the genomes of their bacterial hosts as prophages, it is possible to predict prophage genomes within their bacterial hosts. However, not all predicted prophages can produce virions. Therefore, for the purposes of this study, phage genomes are those that have been experimentally verified to produce virions.

As sequencing capacity has increased, our understanding of the size of phage genomes has also increased. Between 2013 and 2016, a significant number of phages with genomes >200 kb were sequenced and dubbed ''jumbo phages,''[16] although the isolation of ''jumbo phages'' is still thought to be rare. More recently, phages with genomes >500 kb have been reconstructed from metagenomes and referred to as megaphages, further expanding the known size of phage genomes.[17]

The greater number of phage genomes available results in common analyses, including (1) comparative genomic analyses,[18,19] (2) taxonomic classification,[20–23] (3) software for prediction of novel phages,[24–29] and (4) often the first step in analysis of viromes is the comparison of sequences with a known database. The huge amount of potential resource within phage genomes requires a comprehensive set of complete and consistently curated genomes from cultured isolates that can be used to build databases for further analyses.

When analyzing new phage genomes, it is important to know exactly how many phage genomes you are comparing the search with, and any biases (or not) inherent in that data set. Although this should be a relatively trivial question to answer, it is not because there are currently no such databases that contain only complete phage genomes that allow extraction in an automated reproducible manner. Although RefSeq provides well annotated complete phage genomes, it is not representative of the diversity of complete phage genomes. RefSeqs are only created for exemplar phage species, as defined by the International Committee on Taxonomy of Viruses (ICTV). Despite the tremendous work from the ICTV, the process of taxonomy approval is done annually and many phages remain without taxonomy. Thus, RefSeqs will always be catching up with the submission of new phage genomes and lag behind latest submissions. We have created an automated method for researchers to extract complete phage genomes from GenBank in a reproducible manner for use in genomic and metagenomic analyses, and provide general properties of the data set, thus allowing for better understanding of its features and limitations.

Materials and Methods

Phage genomes were download using the ''PHG'' identifier along with minimum and maximum length cutoffs. We also assume the genomes are from phages that have been shown to produce virions and are not predictions of prophages, a requirement of submitting phage genomes. Genomes were filtered based on several parameters to identify complete and near complete phage genomes. This includes initial searching for the term ''Complete'' and ''Genome'' in the phage description, followed by ''Complete'' and (''Genome'' or ''Sequence'') or a genome length of >10 kb. The list of genomes was then manually curated to identify obviously incomplete phage genomes, the accession numbers of genomes that are obviously incomplete were added to an exclusion list. As new genomes are added to GenBank continually, the INfrastructure for a PHAge REference Database (INPHARED) is designed to be updated continually. The use of an exclusion list allows the same incomplete genomes to be identified each time it is updated. An exclusion list is maintained on GitHub that can be added by the community. Although this process is not perfect, it provides a mechanism for the community to manually curate complete phage genomes that is better than one individual checking thousands of genomes repeatedly. Efforts to identify ''false hits'' were reported by many researchers, we would like to thank all members of the phage community who helped in initial curation.

After filtering, genes are called using Prokka with the—noanno flag, with a small number of phages using—gcode 15.[17,30] Gene calling was repeated to provide consistency across all genomes, which is essential for comparative genomics. A prebuilt database (https://doi.org/10.25392/leicester.data.14242085) is provided so gene calling only occurs on newly deposited genomes. The original GenBank files are used to gather metadata including taxa and bacterial host, and the Prokka output files are used to gather data relating to genomic features. The gathered data are summarized in a tab-delimited file that includes the following: accession number, description of the phage genome, GenBank classification, genome length (bp), molecular GC (%), modification date, number of coding sequences (CDS), proportion of CDS on positive sense strand (%), proportion of CDS on negative sense strand (%), coding capacity (%), number of transfer-RNAs (tRNAs), bacterial host, viral genus, viral subfamily, viral family, viral realm, Baltimore group (derived from phylum), and the lowest viral taxa available (from genus, subfamily, and family). Coding capacity was calculated by comparing the genome length with the sum length of all coding features within the Prokka output, and tRNAs were identified by the use of tRNA identifier. Other outputs include a fasta file of all phage genomes, a MASH index for rapid comparison of new sequences, vConTACT2 input files, and various annotation files for IToL and vConTACT2. The vConTACT2 input files produced from the script were processed using vConTACT2 v0.9.13 with—rel-mode Diamond—db ''None''—pcs-mode MCL—vcs-mode ClusterONE—min-size 1 and the resultant network was visualized using Cytoscape v3.8.0.[31,32]

To identify genes indicative of a temperate lifestyle within genomes, we used a set of protein families Hidden Markov Models (HMM) as described previously.[33,34] These HMMs cover the integrase and transposase genes that are associated with the known integration methods of phages into bacterial genomes (PF07508, PF00589, PF01609, PF03184, PF02914, PF01797, PF04986, PF00665, PF07825, PF00239, PF13009, PF16795, PF01526, PF03400, PF01610, PF03050, PF04693, PF07592, PF12762, PF13359, PF13586, PF13610, PF13612, PF13701, PF13737, PF13751, PF13808, PF13843, and PF13358).[33,34] If a genome encoded one of these genes, it was assumed to be temperate. Antimicrobial resistance genes (ARGs) and virulence factors were identified using Abricate with the resfinder and VFDB databases using 95% identity and 75% coverage cutoffs.[35–37]

The phylogeny of ''jumbo phages'' was constructed from the amino acid sequence of the TerL protein, extracted from 313/314 of the ''jumbo phage'' genomes. Sequences were queried against a database of proteins from non''jumbo phages'' using Blastp and the top 5 hits were extracted[38] with redundant sequences being removed. Sequences were aligned with MAFFT, with a phylogenetic tree being produced using IQ-Tree with ''-m WAG -bb 1000'' that was visualized using IToL.[39–41] Additional information was overlaid using IToL templates that are generated through INPHARED.

Rarefaction analysis was carried out for phage genomes from the top 10 most common hosts. Phage genomes were clustered at the level of genus if they belonged to the same vConTACT2 subcluster, and species using ClusterGenomes v5.1 (95% ID over 95% length)[42] on the final set of non-deduplicated genomes, although RefSeq duplicates had been removed at this point. An additional set of these genomes pooled together was included. Rarefaction curves and species richness estimates were produced using Vegan in R.[43,44]

All data from January 2021 are available at Figshare https://doi.org/10.25392/leicester.data.14242085 and the script used for downloading and analyzing genomes is available on GitHub (https://github.com/RyanCook94/).

Results

The output of the INPHARED script provides a set of complete phage genomes, where all genes have been called in a consistent manner that allows comparative genomics and phylogenetic analysis. Unlike RefSeq, it will include all complete phage genomes, including those that have not been classified by the ICTV, and strains of the same phage species (or genome neighbors as they are referred to in the National Center for Biotechnology Information [NCBI] Viral Genomes Resource). In addition, it provides a MASH database to allow rapid comparison of new phage genomes against to identify close relatives, along with formatted databases for input into vConTACT2 to allow identification of more distant relatives. The host data (genus) for each phage are extracted along with summary information for each genome, which is reformatted to allow overlay onto trees in IToL so that the most common analyses for classification of new phages can be easily produced (see Supplementary Fig. S1 for full details).

For this study, we used a lenient definition of ''complete'' to identify complete phage genomes. Strictly speaking, a complete phage genome would include the terminal ends of the phage genome, but because many phages are sequenced using a transposon-based library preparation,[15,19] these terminal bases are never obtained (as transposons have to insert between bases). Another limitation for completeness is that for phage genomes with long terminal repeats; if the length of the repeat is larger than the library insert size, the repeats cannot be resolved. Details of library preparation, and if terminal ends have been confirmed, are not included in GenBank files, thus preventing automated retrieval of this information.

We then identify how many phage genomes have been sequenced to date and 18,134 genomes were extracted from GenBank. Of these, 3890 phage genomes are RefSeq entries that are derived from primary submissions, resulting in 14,244 complete phage genomes.

Current recommendations by the ICTV are that phages are uniquely named.[45] Assuming a unique name represents a unique phage there are 12,127 phages. However, there are multiple examples of phages with the same name that are not genetically identical. Thus, phage names are not a suitable method for determining the number of unique phage genomes. As an alternative, deduplication of genomes at 100%, 97%, and 95% identity results in 13,830, 12,845, and 12,770 genomes, respectively.

Having established a data set of ''complete'' phage genomes, we then analyzed these data to look at how the number of phage genomes being sequenced over time is changing, the host they are isolated on, and overall genomic properties. First, we looked at the increase in the number of phage genomes that are sequenced over time. Although the number of phage genomes has rapidly increased over the past 20 years, the rate of increase has slowed in the past decade (Fig. 1), with the number of phage genomes doubling every 2–3 years.

Phage hosts and predicted gene function

Utilizing our INPHARED, database we extracted the bacterial hosts and information on the predicted number of

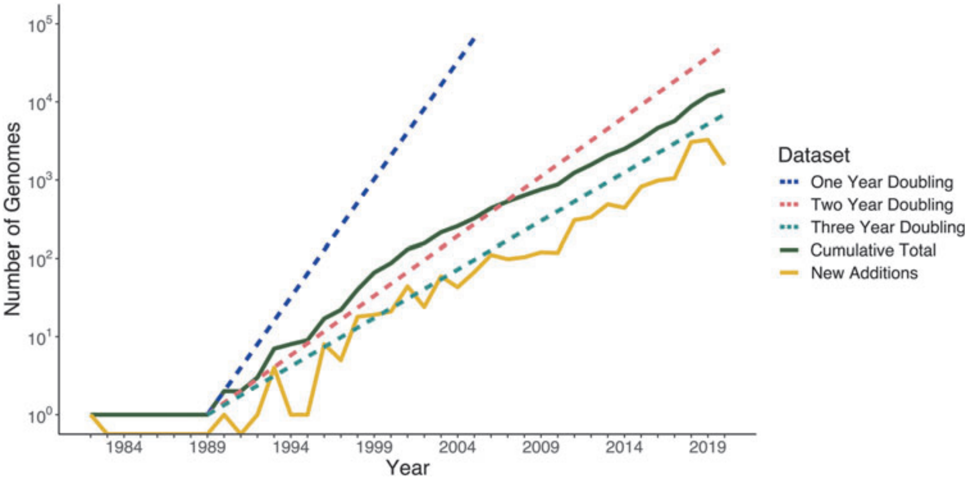


FIG. 1. Number of complete phage genomes in GenBank over time. Dates were estimated based on date of submission (for 235 genomes, the date of update was used as no submission date was available). The reference lines showing doubling rates (dashed) begin in 1989, as this is when the number of phage genomes increased beyond the first submission in 1982.

''hypothetical'' proteins for each phage, so those with no predicted function. Across all phages, 56% of genes encoded hypothetical proteins, supporting the often quoted idea that the majority of genes encode proteins within unknown function.[46]

The host of 87% (12,402/14,244) of phages could be identified, with 13% of phages not having a known host or identifiable host information in the GenBank file, resulting in the genomes of phages infecting 234 different hosts (bacterial genera) having been sequenced. However, there is a clear bias in the isolation of phages against the same host (Fig. 2A). Phages that infect *Mycobacterium* spp. are the most commonly deposited genomes (~13%), largely due to the pioneering work of the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program.[47] This is followed by *Escherichia* spp., *Streptococcus* spp., and *Pseudomonas* spp. (Fig. 2A).

Phages isolated on just 30 different bacterial genera account for ~75% of all phage genomes in the database (Supplementary Table S1). For nondeduplicated genomes isolated against the top 10 hosts specified in the GenBank file, we used rarefaction analysis to determine the diversity of these genomes and establish redundancy with respect to host. Using a cutoff of 95% identity over 95% length to define a species and vConTACT2 subclusters to define a genus, the number of phages continues to increase with the number of genomes sequenced (Fig. 3). Suggesting that there is little redundancy within the database and we are not reaching the point where identifying new phage species is a rare event. Utilizing the rarefaction data for the top 10 hosts, we estimated how many different species of phages might infect each of these different bacterial genera (Supplementary Table S2). For *Mycobacterium*, there are 695 current phage species that lead to an estimation of 2132–2282 species that might infect *Mycobacterium*. Thus, even for hosts wherein thousands of phages have been isolated, we are only just scratching the surface of total phage diversity. We are also likely underestimating the total number of different phage species. In the case of *Mycobacterium*, a large proportion of phages have been isolated on only a single strain as part of the SEA-PHAGES program.[47] Thus, these phages are unlikely to be representative of phages that infect all bacterial species within the genus *Mycobacterium*. Increasing the diversity of the host *Mycobacterium*, that is, using more species of *Mycobacterium* for phage isolation, is likely to lead to more species of phage being isolated, increasing our estimates.

Lytic and temperate phages

To identify whether phages are lytic or temperate, we searched for genes that facilitate a temperate lifestyle (e.g., integrase and recombinase) that have been used in previous studies to predict lytic/temperate phages.[33,34] This process is only a prediction and having such genes does not always mean the phage will enter a lysogenic cycle. However, it is a useful starting point that facilitates large scale comparative analyses when experimental data for all phages are either not available or readily accessible on such a scale.

Within the INPHARED data set, 4258 (~30%) phages have the potential to access a lysogenic lifecycle. The frequency of putative temperate phages was highly variable depending on the host (Supplementary Fig. S2). The number of putative temperate phages is also biased toward a small number of hosts with 1217, 846, and 214 isolated on *Mycobacterium*, *Streptococcus*, and *Gordonia*, respectively. Collectively, these three hosts account for ~54% of all putative temperate phage genomes sequenced to date (Supplementary Fig. S2).

Genomic properties

Genome sizes. Phage genomes ranged from 3.1 to 642.4 kb in size, with a wide distribution in the size of genomes with several observable peaks in genome size. The most prominent peaks are at 5–10, 40, 50, and ~165 kb (Fig. 2B).

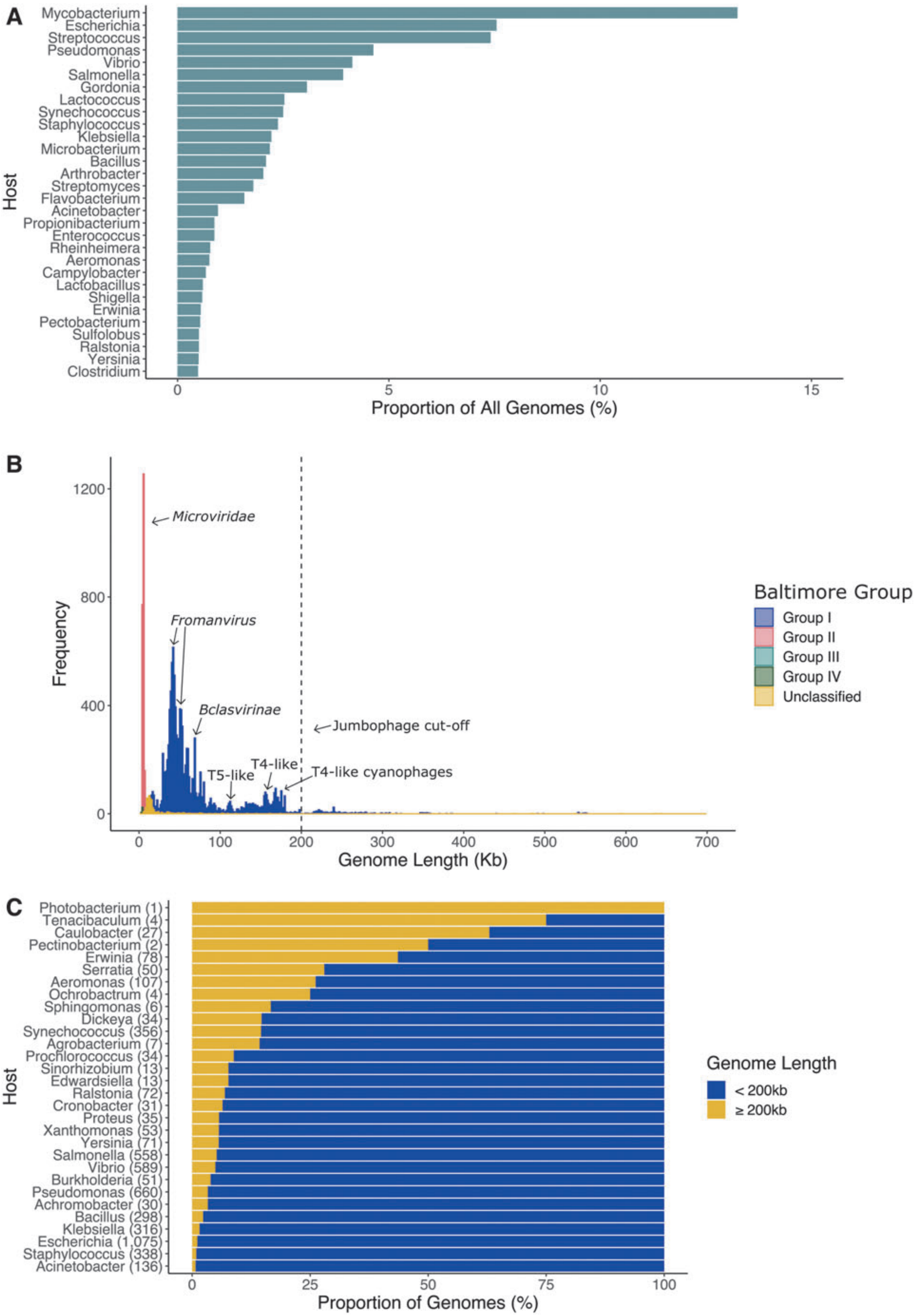
Coding capacity. The mean and median coding capacity was 90.45% and 91.52%, respectively (Supplementary Fig. S2). Of the 14,244 genomes, 5731 (~40%) have ≥90% of coding features on one strand and 3293 (~23%) of these are entirely on one strand (Supplementary Fig. S2). The number of phages with genes encoding tRNAs was 4590 (~32%) and the number of tRNAs ranged from 1 to 62 with a median of 3 (mean of 7.23, and mode of 1). Although there is much literature on phage-encoded tRNAs, the roles they play remain unclear.[48]

Jumbo phages. Phages with genomes >200 kb are often referred to as ''jumbo phages'' and are reported to be ''rarely isolated''[16] and indeed only 314 genomes (~2.2%) fitting this definition were identified, suggesting that they are indeed rare. To further investigate whether ''jumbo phages'' are as rare as is thought, we looked at the distribution in the context of the previously identified host bias. ''Jumbo phages'' have only been isolated on 31 of 234 identifiable bacterial hosts (Supplementary Table S1) and are far more commonly isolated on some hosts than others. Noticeably absent are any ''jumbo phages'' that infect *Mycobacterium*, *Gordonia*, *Lactococcus*, *Arthrobacter*, and *Streptococcus*, with >4000 phages having been sequenced from these bacterial hosts (Fig. 2C).

For host bacteria that have had far fewer phages isolated on them such as *Caluobacter*, *Sphingomonas*, *Erwinia*, *Areomonas*, *Dickeya*, and *Ralstonia*, the frequency of ''jumbo phage'' isolation is far higher (Fig. 2C). Owing to the small

→

FIG. 2. Overall properties of phages. **(A)** Proportion of phages isolated on the top 30 most abundant hosts. **(B)** Distribution of phage genome sizes with colors indicating Baltimore group and labels indicating typical phages for prominent peaks. **(C)** Proportion of ''jumbo phages'' on top 30 hosts for which at least one ''jumbo phage'' has been isolated with the total number of phages isolated against that host shown in brackets after its name.


A
Host
Mycobacterium
Escherichia
Streptococcus
Pseudomonas
Vibrio
Salmonella
Gordonia
Lactococcus
Synechococcus
Staphylococcus
Klebsiella
Microbacterium
Bacillus
Arthrobacter
Streptomyces
Flavobacterium
Acinetobacter
Propionibacterium
Enterococcus
Rheinheimera
Aeromonas
Campylobacter
Lactobacillus
Shigella
Erwinia
Pectobacterium
Sulfolobus
Ralstonia
Yersinia
Clostridium
0
5
10
15
Proportion of All Genomes (%)
B
Frequency
1200
800
400
0
Microviridae
Fromanvirus
Bclasvirinae
Jumbophage cut-off
T5-like
T4-like
T4-like cyanophages
0
100
200
300
400
500
600
700
Genome Length (Kb)
Baltimore Group
Group I
Group II
Group III
Group IV
Unclassified
C
Host
Photobacterium (1)
Tenacibaculum (4)
Caulobacter (27)
Pectinobacterium (2)
Erwinia (78)
Serratia (50)
Aeromonas (107)
Ochrobactrum (4)
Sphingomonas (6)
Dickeya (34)
Synechococcus (356)
Agrobacterium (7)
Prochlorococcus (34)
Sinorhizobium (13)
Edwardsiella (13)
Ralstonia (72)
Cronobacter (31)
Proteus (35)
Xanthomonas (53)
Yersinia (71)
Salmonella (558)
Vibrio (589)
Burkholderia (51)
Pseudomonas (660)
Achromobacter (30)
Bacillus (298)
Klebsiella (316)
Escherichia (1,075)
Staphylococcus (338)
Acinetobacter (136)
0
25
50
75
100
Proportion of Genomes (%)
Genome Length
< 200kb
≥ 200kb

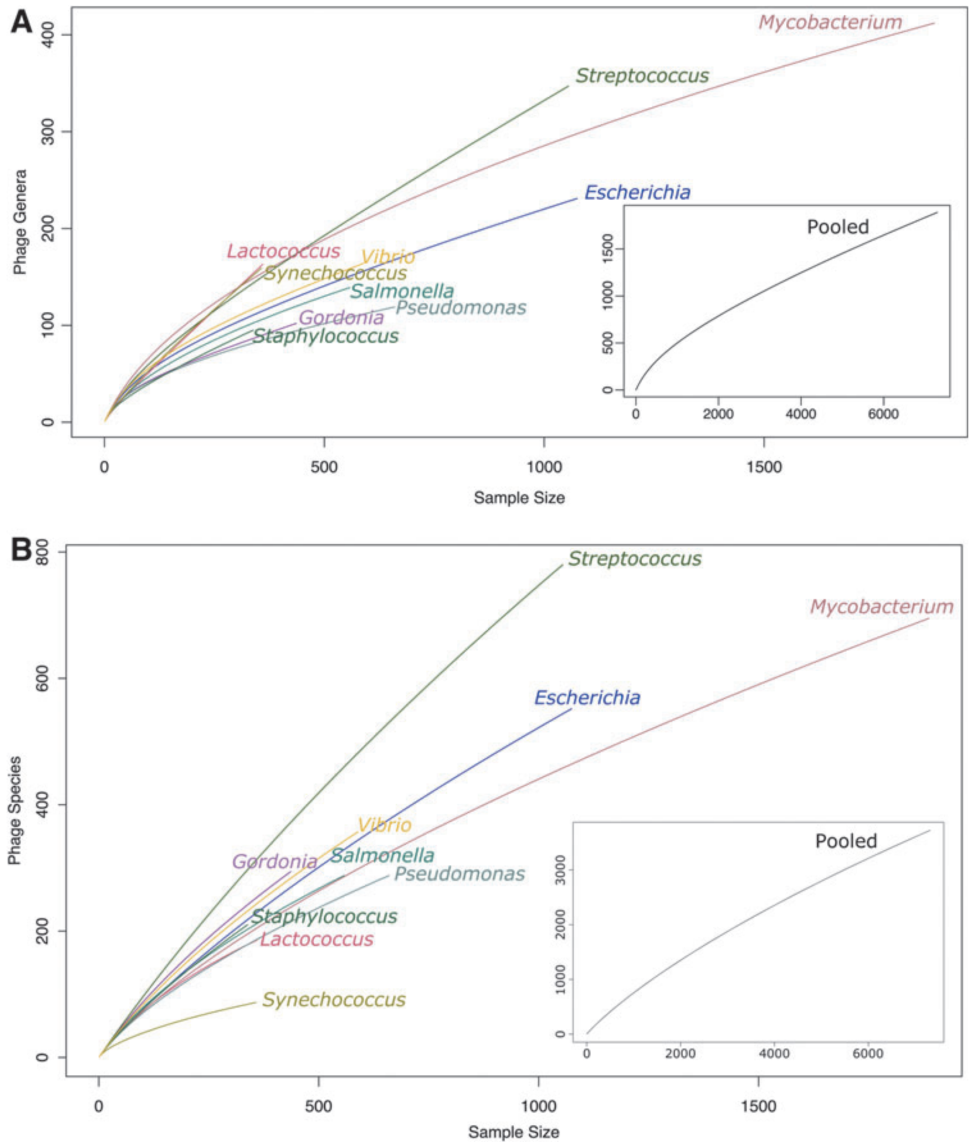


FIG. 3. Genomic diversity of phages on the top 10 most abundant hosts. **(A)** Rarefaction curve of phage genera. Genera were defined by vConTACT2 clustering. **(B)** Rarefaction curve of phage species. Species were defined as 95% identity over 95% of genome length.

sampling depth of some of these hosts (e.g., *Photobacterium* and *Tenacibaclum*), it is not possible to determine whether the high proportion of genomes is merely a result of the low number of genomes sequenced. However, for other hosts such as *Aeromonas*, *Erwinia*, and *Caulobacter* from which >20 phages have been isolated, ∼26%, ∼44%, and ∼63% are categorized as ''jumbo,'' respectively. Therefore suggesting ''jumbo phages'' are not always rare on particular hosts.

We further investigated the phylogeny of ''jumbo phages'' using the translated sequence of the *terL* gene. The ''jumbo phages'' are well distributed across the tree and do not form a single monophyletic clade, suggesting that they have arisen on multiple occasions with 14 clades containing at least one ''jumbo phage.'' Of these 14 clades, 12 also contain a nonjumbo phage. Furthermore, not all ''jumbo phages'' are equal, with ''jumbo'' cyanophages infecting the cyanobacteria *Synechococcus* and *Procholorococcus* only marginally larger than there nonjumbo cyanophages relatives. These ''jumbo phages'' are also more closely related to their nonjumbo cyanophages relatives than other ''jumbo phages'' (Fig. 4). A closer relationship of ''jumbo phages'' with nonjumbo phages than other ''jumbo phages'' is not limited to cyanophages (Fig. 4). A similar pattern of grouping nonjumbo with ''jumbo phages'' is observed when a reticulate approach is used to look at the relatedness of phage genomes using vConTACT2 (Supplementary Fig. S3).

Virulence factors and ARGs. The presence of ARGs and virulence factors is a major concern for phage therapy, as the use of phages carrying such genes may make the populations of bacteria they are intended to kill more virulent or resistant to antibiotics. We therefore, used this database to investigate the frequency and diversity of phage-encoded virulence factors and ARGs. In total, 235 genomes (∼1.6%) were found to encode a putative virulence factor and 43 genomes (∼0.3%) to encode a putative ARG. The most common virulence genes were the stx_{2A} (72 genomes) and stx_{2B} (71 genomes) genes that encode subtypes of the Shiga toxin (Supplementary Table S3). The most common ARGs were the *mef*(A) (14 genomes) and *msr*(D) genes that confer resistance to macrolide antibiotics (Supplementary Table S4).[49] Most genomes encoding a virulence factor were predicted

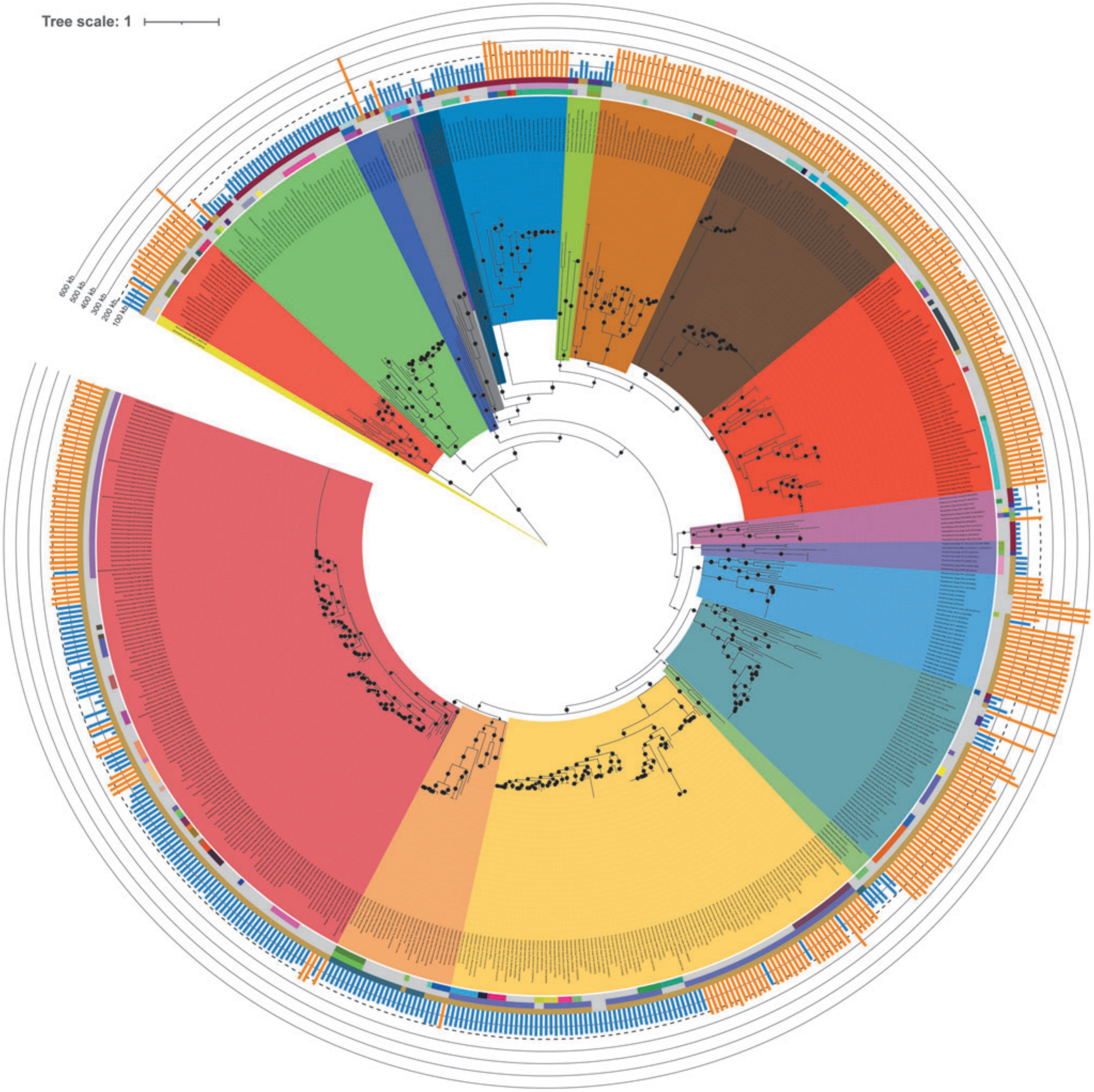


FIG. 4. Phylogenetic tree of translated terL gene for 313 ''jumbo phages'' and their closest relatives. The alignment was produced using MAFFT[39] and tree produced using IQTree using WAG model with 1000 bootstrap repeats.[40] Colored regions indicate viral clades, colored rings indicate viral genus, subfamily, and family (innermost to outermost), and bars indicate genome length with blue and orange bars belonging to nonjumbo and ''jumbo'' phages, respectively. Bootstrap values indicated by black circles are scaled to the bootstrap value, with a minimum value of 70% displayed.

to be from temperate phages (222/235), and were found to infect six bacterial genera, with the three most abundant hosts being *Streptococcus*, *Staphylococcus*, and *Escherichia*, respectively. The hosts for some genomes could not be determined (55/235). The genomes encoding virulence factors were widely distributed over 26 putative genera (Supplementary Fig. S3). All genomes encoding an ARGs were predicted to be temperate and were found to be isolated from eight bacterial genera, with the majority of phages linked to *Streptococcus* spp. (27/43).

Discussion

Defining how many different complete phage genomes have been sequenced is not as simple a question as it might appear. Based on accession numbers, there are 14,244 phage genomes, once RefSeq duplicates have been removed. Using unique names results in 12,127 phages, however, using names alone does not give an accurate estimate of the number of different phages, as genetically different phages have the same name. The use of deduplication at 100% identity suggests 13,830 unique phage genomes (January 2021) from cultured isolates. This also highlights that although RefSeq is a valuable resource, it is not at all representative of phage diversity. INPHARED provides a more comprehensive set of complete phage genomes from cultured phage isolates than RefSeq, in an easily accessible format. There are other resources that provide more comprehensive sets of phage genomes than RefSeq, including the NCBI Viral Genomes Resource.[50,51] The NCBI Viral Genome Resource allows manual filtering of phages through a graphical user interface and access to the same genomes in INPHARED. The automated filtering provided by INPHARED is a key difference, which prevents a user having to exclude the same genomes every time the database is updated. The integrated microbial genomes viral resource (IMG/VR) provides access to >2 million viral genomes, including phages, through a graphical interface.[52] The overwhelming majority of genomes in IMG/VR are constructed from metagenomes and have never been cultured. INPHARED is not designed to replace these valuable resources. The INPHARED provides rapid access to complete phage genomes from cultured phage isolates, without the need for continued manual filtering and provides metadata in an accessible format to allow initial analysis commonly used with phages to be carried out.

The INPHARED reveals clear patterns in phage genomes and biases in the selection of phage genomes that are currently available, but not always discussed in the analysis of genomes. The first is that the number of phage genomes is relatively small. Even for hosts wherein the highest number of phages have been isolated on, our estimates suggest thousands of new phage species remain to be isolated and sequenced. If we consider there are now >300,000 assembled representative bacterial genomes in GenBank, with many hundreds of thousands more for particular genera (e.g., >300,000 *Salmonella* and *Escherichia* genomes alone)[53] compared with only 558 and 1075 of their respective phages, the representation of phage genomes to date is tiny compared with their bacterial hosts. Furthermore, the rate at which phage genomes are being sequenced is slowing down rather than increasing. Given the renewed interest in phages, and increased accessibility of sequencing, the decrease in the rate over time was surprising.

The second point of note is the bias in phage genomes. There is a clear bias in the isolation of phages from a small number hosts, with far more lytic than temperate phages. Thus, these phages are representative of these particular hosts, rather than phages in their entirety. Owing to the enormous success of the SEA-PHAGES program, many phages have been isolated on *Mycobacterium* and *Gordonia*.[54] This in turn results in approximately one-third of all temperate phage genomes being isolated on these two bacterial genera, whereas the remaining two-thirds are distributed across 142 different hosts.

The overrepresentation of phages infecting particular hosts can lead to truisms that may not be correct. For instance, ''jumbo phages,'' those that have genomes >200 kb, are rarely isolated.[16] Analysis of the INPHARED data set suggests ~2.2% of genomes fall into this category. However, this needs to be viewed in the context of the large bias in the hosts used for isolation, with ~75% of phages isolated on only ~16% of bacterial hosts that could be identified. When the number of ''jumbo phages'' is expressed as a percentage of all phage genomes, their isolation is clearly rare. For some hosts, such as *Mycobacterium*, many hundreds of phages isolated on the same host strain have been sequenced without the isolation of a ''jumbo phage,'' suggesting they are truly rare for this host.[47] However, for other hosts such as *Procholorococcus*, *Synechococcus*, *Caulobacter*, and *Erwinia*, the isolation of ''jumbo phages'' is not a rare event. Although methodological adjustments of decreasing agar viscosity and large pore size filters may increase the number of phages isolated that have larger genome sizes,[16] we suggest that using a wider variety of hosts may increase the number of ''jumbo phages'' isolated. Phylogenetic analysis demonstrated that many ''jumbo phages'' are more closely related to nonjumbo phages than other ''jumbo phages.'' Thus, as the number of phage genomes has increased, an arbitrary descriptor of ''jumbo'' for phages with genomes >200 kb in length has less meaning. Recent comparative analysis of 224 ''jumbo phages'' used proteome size and analysis of protein length to determine a cutoff of 180 kb to separate ''jumbo phages'' from other phages. Using a clustering-based approach, three major clades of ''jumbo phages'' were identified.[55] In this study, using *terL* as a phylogenetic marker to determine the phylogeny of 313 ''jumbo phages'' and their closely related phages suggests they have arisen on multiple occasions, as has been demonstrated previously.[55] ''Jumbo phages'' are clearly not monophyletic and what applies to one ''jumbo phage'' does not hold true for many others.[55] As the number and diversity of ''jumbo phages'' increase, the use of the term seems to have less meaning.

With the increasing interest and use of phages for therapy, the isolation of phages that do not contain known virulence factors or ARGs is imperative. How frequently phages encode antibiotic resistance genes is a topic of much debate.[56,57] A previous study of 1181 phage genomes found that they are rarely encoded by phages, with only 13 candidate genes, of which 4 were experimentally tested and found to have no functional antibiotic activity.[46] We estimate that ~0.3% of phage genomes encode a putative ARG (none have been experimentally tested), a finding that is consistent with previous reports of low-level carriage in phage genomes[56] in a data set that is ~10× larger using similarly stringent cutoffs. Critically, all of these ARGs were found in phages that are predicted to be temperate or have been engineered to carry ARGs as a marker for selection. With the frequency of carriage in temperate phages being ~1% overall. However, these data are still biased by the majority of temperate phages being isolated on only three bacterial genera. Notably no ARGs were detected on phages of *Mycobacterium*, which accounts for ~28% of temperate phages. In comparison, ~2.6% (27/1055) of temperate phages of *Streptococcus* carry putative ARGs and 50% of phages from *Erysipelothrix* (1/2) carry putative ARGs. Clearly a much deeper sampling of temperate phages from a broader range of hosts is required to get an accurate understanding of the role of phage in the carriage of ARGs. Based on the skewed data available to date, it seems unlikely there will be issues in the isolation of lytic phages for therapeutic use that carry known ARGs within their genomes. However, we cannot determine whether these lytic phages can spread ARGs through transduction, or through carriage of as-yet uncharacterized ARGs.

Although there is much debate on the presence and importance of ARGs in phage genomes, the role of genes encoding virulence factors is well studied and the process of lysogenic conversion is well known.[7–10] However, how widespread known virulence genes are in phages is not widely reported. We estimate ~1.6% of phages encode at least one putative virulence factor, with the frequency of carriage far higher in temperate phages (5.5%) than in lytic phages (0.13%). Again, these overall percentages are skewed by host bias with no known virulence factors detected in *Mycobacterium* temperate phages (0/1217), in comparison, 72% of temperate phages of *Shigella* (5/7) and 7% (61/846) of *Streptococcus* contain virulence factors. It is currently impossible to determine whether the higher proportion of ARGs and virulence factors in phages of known pathogens is a feature of their biology, or a skew in the database toward phages of clinically relevant isolates.

Given the biases in the data set, it is not clear whether the general phage patterns (e.g., jumbo phages are rarely isolated, more temperate phages on particular hosts, and the carriage of ARGs and virulence genes) are linked to biology or chronic undersampling of phage genomes that results in some bias. We speculate the latter, which distorts some generalizations about phages. Therefore, far deeper sampling of phage genomes across different hosts is required at an increasing rate.

Conclusions

We have provided a method to automate the download of a curated set of complete genomes from cultured phage isolates, providing metadata in a format that can be used as a starting point for many common analyses. Analysis of the current data highlights what we know about phage genomes is skewed by the majority of phages having been isolated from a small number of bacterial genera. Furthermore, the rate at which phage genomes are being deposited is decreasing. Although understanding of genomic diversity is always influenced by the data available, this is particularly acute for phage genomes with so many phages isolated on a small number of hosts. To obtain a greater understanding of phage genomic diversity, larger number of phages, in particular temperate phages, isolated from a broader range of bacteria need to be sequenced.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

R.C. is supported by a scholarship from the Medical Research Foundation National PhD Training Programme in Antimicrobial Resistance Research (MRF-145-0004-TPG-AVISO). A.M., D.J.S., M.J., and J.H. were supported by NERC (NE/N019881/1). A.M. was supported by MRC (MR/T030062/1and MR/L015080/1).

Supplementary Material

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Table S1
Supplementary Table S2
Supplementary Table S3
Supplementary Table S4

References

1. Suttle CA. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol. 2007;5:801–812.
2. Puxty RJ, Millard AD, Evans DJ, et al. Viruses inhibit CO_2 fixation in the most abundant phototrophs on Earth. Curr Biol. 2016;26:1585–1589.
3. Norman JM, Handley SA, Baldridge MT, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160(3):P447–P460.
4. Monaco CL, Gootenberg DB, Zhao G, et al. Altered virome and bacterial microbiome in human immunodeficiency virus-associated acquired immunodeficiency syndrome. Cell Host Microbe. 2016;19(3):P311–P322.
5. Reyes A, Blanton LV., Cao S, et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A. 2015;112(38):11941–11946.
6. Ma Y, You X, Mai G, et al. A human gut phage catalog correlates the gut phageome with type 2 diabetes. Microbiome. 2018;6:24.
7. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science (80-). 1996;272(5270):1910–1914.
8. van Belkum A, Soriaga LB, LaFave MC, et al. Phylogenetic distribution of CRISPR-Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. MBio. 2015;6:e01796-15.
9. Tsao Y-F, Taylor VL, Kala S, et al. Phage morons play an important role in *Pseudomonas aeruginosa* phenotypes. J Bacteriol. 2018;200:e00189-18.
10. O'Brien AD, Newland JW, Miller SF, et al. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. Science (80-). 1984;226(4675):694–696.
11. Breitbart M, Thompson LR, Suttle CA, et al. Exploring the vast diversity of marine viruses. Oceanography. 2007;20: 135–139.
12. Sanger F, Air GM, Barrell BG, et al. Nucliotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265(5596): 687–695.
13. Sepulveda BP, Redgwell T, Rihtman B, et al. Marine phage genomics: The tip of the iceberg. FEMS Microbiol Lett. 2016;363:fnw158.
14. Hatfull GF. Bacteriophage genomics. Curr Opin Microbiol. 2008;11:447–453.
15. Rihtman B, Meaden S, Clokie MRJ, et al. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. PeerJ. 2016;4:e2055.
16. Yuan Y, Gao M. Jumbo bacteriophages: An overview. Front Microbiol. 2017;8:403.
17. Devoto AE, Santini JM, Olm MR, et al. Megaphages infect Prevotella and variants are widespread in gut microbiomes. Nat Microbiol. 2019;4(4):693–700.
18. Javan RR, Ramos-Sevillano E, Akter A, et al. Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis. Nat Commun. 2019;10:4852.
19. Michniewski S, Redgwell T, Grigonyte A, et al. Riding the wave of genomics to investigate aquatic coliphage diversity and activity. Environ Microbiol. 2019;21:2112–2128.
20. Barylski J, Enault F, Dutilh BE, et al. Analysis of Spounaviruses as a case study for the overdue reclassification of tailed phages. Syst Biol. 2020;69(1):110–123.
21. Chibani CM, Farr A, Klama S, et al. Classifying the unclassified: A phage classification method. Viruses. 2019; 11:195.

22. Rohwer F, Edwards R. The phage proteomic tree: A genome-based taxonomy for phage. J Bacteriol. 2002; 184(16):4529–4235.
23. Adriaenssens EM, Wittmann J, Kuhn JH, et al. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol. 2018;163:1125–1129.
24. Ren J, Ahlgren NA, Lu YY, et al. VirFinder: A novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.
25. Ren J, Song K, Deng C, et al. Identifying viruses from metagenomic data by deep learning. Quant Biol. 2020;8(1): 64–77.
26. Roux S, Enault F, Hurwitz BL, et al. VirSorter: Mining viral signal from microbial genomic data. PeerJ. 2015;3: e985.
27. Akhter S, Aziz RK, Edwards R. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res. 2012;40:1–13.
28. Arndt D, Marcu A, Liang Y, et al. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. Brief Bioinform. 2019;;20(4):1560–1567.
29. Bolduc B, Jang H Bin, Doulcier G, et al. vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. PeerJ. 2017;5:e3243.
30. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–2069.
31. Bin Jang H, Bolduc B, Zablocki O, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019; 37:632–639.
32. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–2504.
33. Cook R, Hooton S, Trivedi U, et al. Hybrid assembly of an agricultural slurry virome reveals a diverse and stable community with the potential to alter the metabolism and virulence of veterinary pathogens. Microbiome. 2021;9:65.
34. Clooney AG, Sutton TDS, Shkoporov AN, et al. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe. 2019;26: 764.e5–778.e5.
35. Chen L, Zheng D, Liu B, et al. VFDB 2016: Hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Res. 2016;44:D694–D697.
36. Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67:2640–2644.
37. Seemann T. 2021. Abricate. Github. https://github.com/tseemann/abricate
38. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215:403–410.
39. Nakamura T, Yamada KD, Tomii K, et al. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34:2490–2492.
40. Nguyen L-T, Schmidt HA, von Haeseler A, et al. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.
41. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res. 2019;47:W256–W259.
42. Roux S. 2019. ClusterGenomes. https://github.com/simroux/ClusterGenomes
43. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.
44. Oksanen J, Guillaume Blanchet F, Friendly M, et al. *vegan: Community Ecology Package*. 2020. https://CRAN.R-project.org/package=vegan
45. Adriaenssens EM, Rodney Brister J. How to name and classify your phage: An informal guide. Viruses. 2017;9:1–9.
46. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3(6):504–510.
47. Hatfull GF, Pedulla ML, Jacobs-Sera D, et al. Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. PLoS Genet. 2006;2(6):e92.
48. Bailly-Bechet M, Vergassola M, Rocha E. Causes for the intriguing presence of tRNAs in phages. Genome Res. 2007;17:1486–1495.
49. Daly MM, Doktor S, Flamm R, et al. Characterization and prevalence of MefA, MefE, and the associated msr(D) gene in *Streptococcus pneumoniae* clinical isolates. J Clin Microbiol. 2004;42:3570–3574.
50. Sayers EW, Cavanaugh M, Clark K, et al. GenBank. Nucleic Acids Res. 2020;48:D84.
51. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733.
52. Roux S, Páez-Espino D, Chen I-MA, et al. IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. 2021;49:D764–D775.
53. Zhou Z, Alikhan NF, Mohamed K, et al. The EnteroBase user's guide, with case studies on Salmonella transmissions, *Yersinia pestis* phylogeny, and Escherichia core genomic diversity. Genome Res. 2020;30:138–152.
54. Hanauer DI, Graham MJ, Betancur L, et al. An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. Proc Natl Acad Sci U S A. 2017;114:13531–13536.
55. Iyer LM, Anantharaman V, Krishnan A, et al. Jumbo phages: A comparative genomic overview of core functions and adaptions for biological conflicts. Viruses. 2021;13(1):63.
56. Enault F, Briet A, Bouteille L, et al. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. ISME J. 2017;11:237–247.
57. Debroas D, Siguret C. Viruses as key reservoirs of antibiotic resistance genes in the environment. ISME J. 2019;13: 2856–2867.

Address correspondence to:
Andrew Millard, PhD
Department of Genetics and Genome Biology
University of Leicester
University Road
Leicester, Leicestershire LE1 7RH
United Kingdom

E-mail: adm39@le.ac.uk