

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/159179>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

INfrastructure for a **PHA**ge **RE**ference **D**atabase: Identification of large-scale biases in the current collection of cultured phage genomes

Ryan Cook¹, Nathan Brown², Tamsin Redgwell³, Branko Rihtman⁴, Megan Barnes², Martha Clokie², Dov J. Stekel⁵, Jon Hobman⁵, Michael A. Jones¹, Andrew Millard^{2*}

¹ School of Veterinary Medicine and Science, University of Nottingham, Sutton Bonington Campus, College Road, Loughborough, Leicestershire, LE12 5RD, UK

² Dept Genetics and Genome Biology, University of Leicester, University Road, Leicester, Leicestershire, LE1 7RH, UK

³ COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark

⁴ University of Warwick, School of Life Sciences, Coventry, UK

⁵ School of Biosciences, University of Nottingham, Sutton Bonington Campus, College Road, Loughborough, Leicestershire, LE12 5RD, UK

Corresponding

author: adm39@le.ac.uk

Abstract

Background

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

Materials and Methods

We developed an automated retrieval and analysis system for phage genomes (<https://github.com/RyanCook94/inphared>) to produce the INPHARED database 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 lytic and again varied with host.

Conclusion

Given the bias of currently sequenced phage genomes, we suggest to fully understand phage diversity efforts should be made to isolate and sequence larger numbers 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 via the lysis of their hosts (1), or directly halting the fixation of CO₂ 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 of 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 by causing increased virulence, a notable example being phage CTX which actually encodes the toxins within the genome of *Vibrio cholerae*, that 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 multi-drug 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 utilise parts of their genomes known as auxiliary metabolic genes (AMGs), homologues of host metabolic genes, to modulate their host's metabolism which 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 have 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 work, phage genomes are those that have been experimentally verified to produce virions.

As sequencing has capacity increased, our understanding of the size of phage genomes has also increased. Between 2013-2016, 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 greater than 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: a) comparative genomic analyses (18, 19), b) taxonomic classification (20–23), c) software to prediction for novel phages (24–29), and d) it is often the

first step in analysis of viromes, the comparison of sequences to 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 analysing new phage genomes, it is important to know exactly how many phage genomes you are comparing the search to, and any biases (or not) inherent in that dataset. While this should be a relatively trivial question to answer, it isn't because there are currently no such databases that contain only complete phage genomes that allow extraction in an automated reproducible manner. Whilst 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 phage 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 dataset, 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 cut-offs. 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" & "Genome" in the phage description, followed by "Complete" & ("Genome" or "Sequence") or a genome length of greater than 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 INPHARED database 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 to by the community. Whilst this process is not perfect, it provides a mechanism for the community to manually curate complete phage genomes which 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 summarised in a tab-delimited file that includes the following: accession number, description of the phage genome, GenBank classification, genome length (bp), molGC (%), modification date, number of CDS, proportion of CDS on positive sense strand (%), proportion of CDS on negative sense strand (%), coding capacity (%), number of tRNAs, bacterial host, viral genus, viral sub-family, viral family, viral realm, Baltimore group (derived from phylum), and the lowest viral taxa available (from genus, sub-family, and family). Coding capacity was calculated by comparing the genome length to 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 visualised using Cytoscape v3.8.0 (31, 32).

To identify genes indicative of a temperate lifestyle within genomes, we used a set of PFAM HMMs 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 cut-offs (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 five 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” which was visualised using IToL (39–41). Additional information was overlaid using IToL templates that are generated via INPHARED.

Rarefaction analysis was carried out for phage genomes from the top ten most common hosts. Phage genomes were clustered at the level of genus if they belonged to the same vCONTACT2 sub-cluster (see above), and species using ClusterGenomes v5.1 (95% ID over 95% length) (42) on the final set of non-duplicated 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 is available at Figshare <https://doi.org/10.25392/leicester.data.14242085> and the script used for downloading and analysing 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 neighbours as they are referred to in the 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 is 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 Figure 1 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 (19, 45), 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, 3,890 phage genomes are RefSeq entries which are derived from primary submissions, resulting in 14,244 complete phage genomes.

Current recommendations by the ICTV are that phages are uniquely named (46). Assuming a unique name is 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, de-duplication of genomes at 100%, 97% and 95% identity results in 13,830, 12,845 and 12,770 genomes respectively.

Having established a dataset of “complete” phage genomes, we then analysed this data to look at how the number of phage genomes being sequenced over time is changing, 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. Whilst the number of phage genomes has rapidly increased over the last 20 years, the rate of increase has slowed in the last decade (Figure 1), with the number of phage genomes doubling every 2-3 years.

Phage Hosts and Predicted Gene Function

Utilising our INPHARED database, we extracted the bacterial hosts and information on the predicted number of ‘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 (47).

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 (Figure 2a). Phages that infect *Mycobacterium* spp. are the most commonly deposited genomes (~13%), largely due to the pioneering work of the SEA-PHAGES program (48). This is followed by *Escherichia* spp., *Streptococcus* spp., and *Pseudomonas* spp. (Figure 2a).

Phages isolated on just 30 different bacterial genera account for ~75% of all phage genomes in the database (Supplementary Table 1). For **non-duplicated** genomes isolated against the top ten hosts specified in the GenBank file, we used rarefaction analysis to determine the diversity of these genomes and establish redundancy in respect to host. Using a cut-off of 95% identity **over 95% length** to define a species and **vConTACT2 sub-clusters to define a genus**, the number of phages continues to increase with the number of genomes sequenced (Figure 3). Suggesting that there is little redundancy within the database and we not reaching the point where identifying new phage species is a rare event.

Utilising the rarefaction data for the top ten hosts, we estimated how many different species of phage might infect each of these different bacterial genera (Supplementary Table 4). For *Mycobacterium*, there are 695 current phage species which leads to an estimation of 2132-2282 species that might infect *Mycobacterium*. Thus, even for hosts where 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 (48). 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*, **i.e.**, 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 if 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 is either not available or readily accessible on such a scale.

Within the INPHARED dataset, 4,258 (~30%) phages have the potential to access a lysogenic lifecycle. The frequency of putative temperate phages was highly variable depending on the host (Supplementary Figure 2). The number of putative temperate phages is also biased towards a small number of hosts with 1,217, 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 Figure 2).

Genomic Properties

Genome sizes

Phage genomes ranged from 3.1 kb 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 kb, 40 kb, 50 kb and ~165 kb (Figure 2b).

Coding capacity

The mean and median coding capacity was 90.45% and 91.52%, respectively (Supplementary Figure 2). Of the 14,244 genomes, 5,731 (~40%) have $\geq 90\%$ of coding features on one strand and 3,293 (~23%) of these entirely on one strand (Supplementary Figure 2). The number of phages with genes encoding tRNAs was 4,590 (~32%) and the number of tRNAs ranged from 1 to 62 with a median of 3 (mean of 7.23, and mode of 1). Whilst there is much literature on phage encoded tRNAs the roles they play remain unclear (49).

Jumbo phages

Phages with genomes greater than 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 if “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 1) 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 >4,000 phages having been sequenced from these bacterial hosts (Figure 2c).

For host bacteria that have had far fewer phages isolated on them such as *Caluobacter*, *Sphingomonas*, *Erwinia*, *Aeromonas*, *Dickeya* and *Ralstonia*, the frequency of “jumbo-phage” isolation is far higher (Figure 2c). Due to the small 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 more than 20 phages have been isolated, ~26%, ~44% and ~63% are categorised 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 “jumbo-phage”. Of these 14 clades, 12 also contain a non-jumbo phage. Furthermore, not all “jumbo-phages” are equal, with “jumbo” cyanophages infecting the cyanobacteria *Synechococcus* and *Prochlorococcus* only marginally larger than their non-jumbo cyanophages relatives. These “jumbo-phages” are also more closely related to their non-jumbo cyanophages relatives than other “jumbo-phages” (Figure 4). A closer

relationship of “jumbo-phages” to non-jumbo phages than other “jumbo-phages” is not limited to cyanophages (Figure 4). A similar pattern of grouping non-jumbo with “jumbo-phages” is observed when a reticulate approach is used to look at the relatedness of phage genomes using vConTACT2 (Supplementary Figure 3).

Virulence Factors and Antimicrobial Resistance Genes

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 phag-encoded virulence factors and ARGs. 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 2). The most common ARGs were the *mef(A)* (14 genomes) and *msr(D)* genes which confer resistance to macrolide antibiotics (Supplementary Table 3) (50). Most genomes encoding a virulence factor were predicted 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 virulence factor encoding genomes were widely distributed over 26 putative genera (Supplementary Figure 3). All genomes encoding an ARG 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 de-duplication at 100% identity suggests 13,830 unique phage genomes (January 2021) from cultured isolates. This also highlights that whilst 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 (51, 52). 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 greater than 2 million viral genomes, including phages, via a graphical interface (53). 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 database 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 database 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 the number of phage genomes is relatively small. Even for hosts where 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 more than 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 (54) compared with only 558 and 1,075 of their respective phages, the representation of phage genomes to date is tiny compared to 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. With a clear bias in both the hosts phages are isolated on, and for lytic phages over temperate phages. Thus, these phages are representative of these particular hosts, rather than phages in their entirety. Due to the enormous success of the SEA-PHAGES program, many phages have been isolated on *Mycobacterium* and *Gordonia* (55). This in turn results in $\sim 1/3^{\text{rd}}$ of all temperate phage genomes being isolated on these two bacterial genera, whereas the remaining $2/3^{\text{ds}}$ 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 dataset suggests $\sim 2.2\%$ 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 $\sim 75\%$ of phages isolated on only $\sim 16\%$ 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 (48). However, for other hosts such as *Prochlorococcus*, *Synechococcus*, *Caulobacter*, and *Erwinia*, the isolation of “jumbo-phages” is not a rare event. While 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 many “jumbo-phages” are more closely related to non-jumbo phages than other “jumbo-phages”. Thus, as the number of phage genomes has increased, an arbitrary descriptor of “jumbo” for phages with genomes over 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 cut-off of 180 kb to separate “jumbo-phages” from other phages. Using a clustering-based approach, three major clades of “jumbo-phages” were identified (56). 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 (56). “Jumbo-phages” are clearly not monophyletic and what applies to one “jumbo-phage” does not hold true for many others (56). As the number and diversity of “jumbo-phages” increases, 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 (57, 58). A previous study of 1,181 phage genomes found that they are rarely encoded by phages, with only 13 candidate genes, of which four were experimentally tested and found to have no functional antibiotic activity (47). We estimate ~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 (57) in a dataset that is ~10x larger using similarly stringent cut-offs. 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, this data is 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). 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 via transduction, or through carriage of as-yet uncharacterised ARGs.

Whilst 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 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 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 if 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 towards phages of clinically relevant isolates.

Given the biases in the dataset, it is not clear if 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 under sampling of phage genomes that results in some bias. We speculate the latter, which distorts some generalisations 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. Whilst 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 numbers of phages, in particular temperate phages, isolated from a broader range of bacteria need to be sequenced.

Acknowledgments

Authorship Confirmation Statement

Competing Interests.

Funding

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/T030062/1 and [MR/L015080/1](#))

Figure 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.

Figure 2

Overall properties of phages. A) Proportion of phages isolated on the top 30 most abundant hosts. B) Distribution of phage genome sizes with colours 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 following its name.

Figure 3

Genomic diversity of phages on the top ten 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.

Figure 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). Coloured regions indicate viral clades, coloured rings indicate viral genus, sub-family and family (innermost to outermost), and bars indicate genome length with blue and orange bars belonging to non-jumbo and “jumbo” phages respectively. Bootstrap values indicated by black circles are scaled to the bootstrap value, with a minimum value of 70% displayed.

References

1. Suttle CA. 2007. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* 5:801–12.
2. Puxty RJ, Millard AD, Evans DJ, Scanlan DJ. 2016. Viruses inhibit CO₂ fixation in the most abundant phototrophs on Earth. *Curr Biol* 26:1585–1589.
3. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* <https://doi.org/10.1016/j.cell.2015.01.002>.
4. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge MT, Wilen CB, Flagg M, Norman JM, Keller BC, Luévano JM, Wang D, Boum Y, Martin JN, Hunt PW, Bangsberg DR, Siedner MJ, Kwon DS, Virgin HW. 2016. Altered Virome and Bacterial Microbiome in Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. *Cell Host Microbe* <https://doi.org/10.1016/j.chom.2016.02.011>.
5. Reyes A, Blanton L V., Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW, Rohwer F, Gordon JI. 2015. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A* <https://doi.org/10.1073/pnas.1514285112>.
6. Ma Y, You X, Mai G, Tokuyasu T, Liu C. 2018. A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome* <https://doi.org/10.1186/s40168-018-0410-y>.

7. Waldor MK, Mekalanos JJ. 1996. Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science* (80-) <https://doi.org/10.1126/science.272.5270.1910>.
8. van Belkum A, Soriaga LB, LaFave MC, Akella S, Veyrieras J-B, Barbu EM, Shortridge D, Blanc B, Hannum G, Zambardi G, Miller K, Enright MC, Mugnier N, Brami D, Schicklin S, Felderman M, Schwartz AS, Richardson TH, Peterson TC, Hubby B, Cady KC. 2015. Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *MBio* 6:e01796-15.
9. Tsao Y-F, Taylor VL, Kala S, Bondy-Denomy J, Khan AN, Bona D, Cattoir V, Lory S, Davidson AR, Maxwell KL. 2018. Phage Morons Play an Important Role in *Pseudomonas aeruginosa* Phenotypes. *J Bacteriol* 200:e00189-18.
10. O'Brien AD, Newland JW, Miller SF, Holmes RK, Smith HW, Formal SB. 1984. Shiga-like toxin-converting phages from *Escherichia coli* strains that cause hemorrhagic colitis or infantile diarrhea. *Science* (80-) <https://doi.org/10.1126/science.6387911>.
11. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. 2007. Exploring the vast diversity of marine viruses. *Oceanography* 20:135–139.
12. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*.
13. Perez Sepulveda B, Redgwell T, Rihtman B, Pitt F, Scanlan DJ, Millard A. 2016. Marine phage genomics: the tip of the iceberg. *FEMS Microbiol Lett* 363:fnw158.
14. Hatfull GF. 2008. Bacteriophage genomics. *Curr Opin Microbiol* 2008/10/14. 11:447–453.
15. Rihtman B, Clokie MRJ, Koskella B, Millard AD. 2016. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. *PeerJ*.
16. Yuan Y, Gao M. 2017. Jumbo bacteriophages: An overview. *Front Microbiol* <https://doi.org/10.3389/fmicb.2017.00403>.
17. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, Archie EA, Turnbaugh PJ, Seed KD, Blehman R, Aarestrup FM, Thomas BC, Banfield JF. 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat Microbiol* <https://doi.org/10.1038/s41564-018-0338-9>.

18. Javan RR, Ramos-Sevillano E, Akter A, Brown J, Brueggemann A. 2018. Prophages and satellite prophages are widespread among *Streptococcus* species and may play a role in pneumococcal pathogenesis. bioRxiv 502740.
19. Michniewski S, Redgwell T, Grigonyte A, Rihtman B, Aguilo-Ferretjans M, Christie-Oleza J, Jameson E, Scanlan DJ, Millard AD. 2019. Riding the wave of genomics to investigate aquatic coliphage diversity and activity. Environ Microbiol 2019/04/04. 21:2112–2128.
20. Barylski J, Enault F, Dutilh BE, Schuller MB, Edwards RA, Gillis A, Klumpp J, Knezevic P, Krupovic M, Kuhn JH, Lavigne R, Oksanen HM, Sullivan MB, Jang H Bin, Simmonds P, Aiewsakun P, Wittmann J, Tolstoy I, Brister JR, Kropinski AM, Adriaenssens EM. 2019. Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages. Syst Biol 0:1–14.
21. Chibani CM, Farr A, Klama S, Dietrich S, Liesegang H. 2019. Classifying the Unclassified: A Phage Classification Method. Viruses 11:195.
22. Rohwer F, Edwards R. 2002. The phage proteomic tree: A genome-based taxonomy for phage. J Bacteriol <https://doi.org/10.1128/JB.184.16.4529-4535.2002>.
23. Adriaenssens EM, Wittmann J, Kuhn JH, Turner D, Sullivan MB, Dutilh BE, Jang H Bin, van Zyl LJ, Klumpp J, Lobočka M, Moreno Switt AI, Rumnieks J, Edwards RA, Uchiyama J, Alfenas-Zerbini P, Petty NK, Kropinski AM, Barylski J, Gillis A, Clokie MRC, Prangishvili D, Lavigne R, Aziz RK, Duffy S, Krupovic M, Poranen MM, Knezevic P, Enault F, Tong Y, Oksanen HM, Rodney Brister J. 2018. Taxonomy of prokaryotic viruses: 2017 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol 163:1125–1129.
24. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5:69.
25. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Sun F. 2018. Identifying viruses from metagenomic data by deep learning.
26. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985.
27. Akhter S, Aziz RK, Edwards R a. 2012. PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. Nucleic Acids Res 40:1–13.

28. Arndt D, Marcu A, Liang Y, Wishart DS. 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Brief Bioinform* 1–8.
29. Bolduc B, Jang H Bin, Doucier G, You Z-Q, Roux S, Sullivan MB. 2017. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* 5:e3243.
30. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069.
31. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 37:632–639.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.
33. Cook R, Hooton S, Trivedi U, King L, Dodd CER, Hobman JL, Stekel DJ, Jones MA, Millard AD. 2021. 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* 9:65.
34. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, Hill C. 2019. Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* 26:764-778.e5.
35. Chen L, Zheng D, Liu B, Yang J, Jin Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res* 2015/11/17. 44:D694–D697.
36. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 2012/07/10. 67:2640–2644.
37. Seemann T. Abricate. Github.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410.
39. Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* 34:2490–2492.
40. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2014/11/03. 32:268–274.

41. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47:W256–W259.
42. GitHub - simroux/ClusterGenomes: Archive for ClusterGenomes scripts.
43. Team RC. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
44. Oksanen J, Guillaume Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens M henry H, Szoecs E, Wagner H. 2020. *vegan*: Community Ecology Package.
45. Rihtman B, Meaden S, Clokie MRJ, Koskella B, Millard AD, Rihtman B Clokie MRJ, Koskella B, Millard AD. MS. 2016. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. *PeerJ* 4:e2055.
46. Adriaenssens EM, Rodney Brister J. 2017. How to name and classify your phage: An informal guide. *Viruses* 9:1–9.
47. Edwards RA, Rohwer F. 2005. Viral metagenomics. *Nat Rev Microbiol* 3:504–510.
48. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM, Hryckowian AJ, Kelchner VA, Namburi S, Pajcini K V., Popovich MG, Schleicher DT, Simanek BZ, Smith AL, Zdanowicz GM, Kumar V, Peebles CL, Jacobs WR, Lawrence JG, Hendrix RW. 2006. Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genet* <https://doi.org/10.1371/journal.pgen.0020092>.
49. Bailly-Bechet M, Vergassola M, Rocha E. 2007. Causes for the intriguing presence of tRNAs in phages. *Genome Res*2007/09/04. 17:1486–1495.
50. Daly MM, Doktor S, Flamm R, Shortridge D. 2004. Characterization and prevalence of MefA, MefE, and the associated msr(D) gene in *Streptococcus pneumoniae* clinical isolates. *J Clin Microbiol* 42:3570–3574.
51. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2020. GenBank. *Nucleic Acids Res* 48:D84.
52. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-

- Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44:D733.
53. Roux S, Páez-Espino D, Chen I-MA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. 2021. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res* 49:D764–D775.
54. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Achtman M. 2020. The Enterobase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res* 30:138–152.
55. Hanauer DI, Graham MJ, Betancur L, Bobrownicki A, Cresawn SG, Garlena RA, Jacobs-Sera D, Kaufmann N, Pope WH, Russell DA, Jacobs WR, Sivanathan V, Asai DJ, Hatfull GF. 2017. 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* 114:13531–13536.
56. M. Iyer L, Anantharaman V, Krishnan A, Burroughs AM, Aravind L. 2021. Jumbo Phages: A Comparative Genomic Overview of Core Functions and Adaptions for Biological Conflicts. *Viruses* .
57. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. *ISME J* 11:237–247.
58. Debroas D, Siguret C. 2019. Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J* 13:2856–2867.