Mobile DNA

REVIEW Open Access

# Viral communities of the human gut: metagenomic analysis of composition and dynamics

Varun Aggarwala[1†], Guanxiang Liang[1,2†] and Frederic D. Bushman[1*]

## Abstract

**Background:** The numerically most abundant biological entities on Earth are viruses. Vast populations prey on the cellular microbiota in all habitats, including the human gut.

**Main body:** Here we review approaches for studying the human virome, and some recent results on movement of viral sequences between bacterial cells and eukaryotic hosts. We first overview biochemical and bioinformatic methods, emphasizing that specific choices in the methods used can have strong effects on the results obtained. We then review studies characterizing the virome of the healthy human gut, which reveal that most of the viruses detected are typically uncharacterized phage - the viral dark matter - and that viruses that infect human cells are encountered only rarely. We then review movement of phage between bacterial cells during antibiotic treatment. Here a radical proposal for extensive movement of antibiotic genes on phage has been challenged by a careful reanalysis of the metagenomic annotation methods used. We then review two recent studies of movement of whole phage communities between human individuals during fecal microbial transplantation, which emphasize the possible role of lysogeny in dispersal.

**Short conclusion:** Methods for studying the human gut virome are improving, yielding interesting data on movement of phage genes between cells and mammalian host organisms. However, viral populations are vast, and studies of their composition and function are just beginning.

**Keywords:** Virus, Bacteriophage, Virome, Microbiome, Metagenomics, DNA, Transduction

## Background

The human virome is overwhelmingly composed of unstudied bacterial viruses of unknown importance to health and disease. Here we overview metagenomic methods for studying these populations, and some recent results.

## Main text

### Introduction

Global viral populations are vast. Rich sea water typically harbors $10^6$ bacterial cells per ml, but virus-like particles (VLPs) outnumber cells by a factor of ten [1–3]. Given the enormous number of VLPs, it is generally impossible to determine how many really correspond to infectious viruses. However, Electron microscope (EM) analysis shows that many have morphologies resembling bacterial viruses [2, 3], so it seems likely that most VLPs are real viruses. The viral populations living in healthy humans are also enormous. The human microbiome contains roughly 100 trillion cells, equaling or exceeding the number of human cells comprising our bodies [4]. Stool from healthy individuals can contain ~$10^{11}$ cells per gram, which are predominantly bacteria, but also contain archaea and microeukaryotes [5–9]. Studies are just beginning on the viral populations associated with our microbiota, but early work has established that the communities are large and dynamic [10–19].

Here we review recent studies of the human virome. Several excellent reviews have summarized a variety of aspects (e. g. [11, 20–24])—here we first review techniques for purifying viral particles, emphasizing that different methods yield different parts of the viral population. We then review bioinformatic pipelines for analyzing the output, focusing on strengths and weakness of

* Correspondence: bushman@mail.med.upenn.edu
†Equal contributors
[1]Department of Microbiology, University of Pennsylvania School of Medicine, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA
Full list of author information is available at the end of the article

current technology. We particularly emphasize the challenges posed by the "viral dark matter" [11, 25] —in metagenomic studies of the human virome, the vast majority of reads cannot be annotated into functional or taxonomic categories (Fig. 1). This is likely because of the enormous size and diversity of global viral populations, and the fact that only a few thousand viral genomes (7321 from NCBI Genome) are available in databases, so that any new virus captured from nature will usually not have much resemblance to a database entry. Following the review of methods, we summarize a few recent studies that illuminate the nature of the human gut virome and transfer of phage DNA sequences between cells and between humans.

## Biochemical methods for purifying and sequencing VLP genomes

It is possible to study the viral populations of human gut by purifying DNA from total stool, then sequencing and aligning the reads to viral databases [26]. However, viral DNA represents only a small minority of the total DNA recovered, and most viral sequences do not closely resemble viral genomes available in databases (the dark matter problem mentioned above) [10, 11, 15]. To provide a more comprehensive picture, it is often useful to isolate VLPs first from the sample, and then analyze the viral metagenome de novo in the sample of interest [27].

The methods used for viral particle purification have a strong effect on the populations recovered. An investigator must decide whether they want to study viral genomes made of DNA, RNA or both, and whether they want to study both enveloped and non-enveloped viruses.

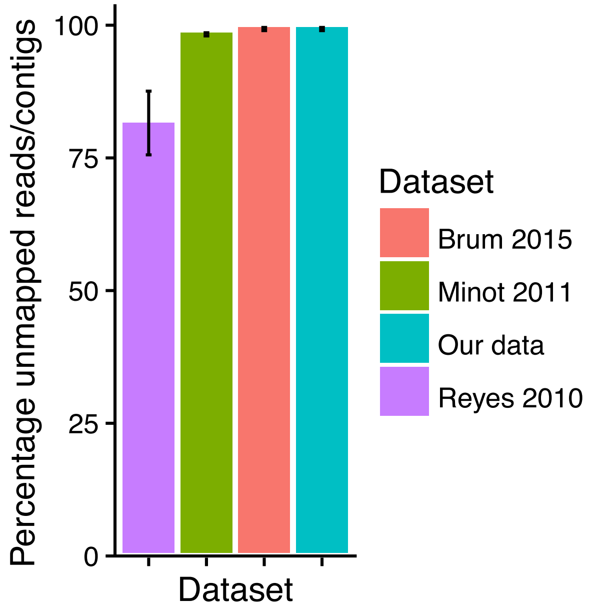

**Fig. 1** Illustration of the viral dark matter problem. Percentage of unmapped reads or contigs in several viral purified sequencing studies and on 849 viral purified sequencing datasets collected locally at University of Pennsylvania

In a typical protocol, feces are suspended in a buffer, and then filtration or centrifugation steps are added to remove bacterial or human cells and any particulate material [27]. Protocols vary in the amount of starting material required (0.1 g to 5 g) [10, 12–15, 28], buffers used (saline-magnesium (SM) buffer [10, 13–15]; phosphate-buffered saline (PBS) buffer [17, 29], and filter pore size. Commonly used are 0.2 and 0.45 μm, but some phages and eukaryotic viruses are larger than 0.2 μm [30]. Going the other way, bacteria smaller than 0.45 μm have been reported, so the larger pore size may result in sporadic bacterial contamination [30]. Following filtration, protein purification filters, such as Centricon Plus-70 Centrifugal Filter (Millipore) are often used for further purifying and concentrating VLPs [31]. As an alternative, cesium chloride (CsCl) density gradient centrifugation, can be used for further VLPs purification and enrichment [14, 15]. A recent study reported that including a CsCl density gradient step was better than other methods in removing host-derived DNA [30]. However, this method is time intensive, which limits the number of samples that can be processed in parallel [30].

Chloroform can be added to disrupt the cell membrane, allowing further removal of microbial and host cells and debris [14, 15, 17]. However, a disadvantage is that enveloped viruses will also be removed, and there may be other effects on viral populations as well. Thus, some researchers choose not to treat VLP preps with chloroform. This allows a more comprehensive assessment of the viruses present, but also results in more contamination with nucleic acids from cells and cellular debris, usually meaning that downstream bioinformatic steps must be relied on to distinguish viral sequences from background. The differences among methods are summarized in Table 1.

After VLPs are isolated, free nucleic acids are removed by treating VLPs with DNase and RNase. The viral DNAs and RNAs can then be extracted by any of several methods, including standard phenol-chloroform methods [10, 12], Trizol-based methods [32], or commercial kits, such as DNeasy (Qiagen) [13, 15], or QIAmp Ultrasens Virus kit (Qiagen) [33].

The yield of nucleic acids extracted from VLPs is usually low, necessitating an amplification step before sequence analysis. A common method for DNA samples is multiple displacement amplification (MDA), which uses the highly processive phage phi29 DNA polymerase primed with random oligonucleotides to amplify viral genomes. A disadvantage of MDA is that it will preferentially amplify small circular viruses by rolling circle amplification [34]. For analyzing RNA viruses, VLP RNA

**Table 1** Methods for purifying VLPs

| VLPs isolation steps | Methods | Pros | Cons | References |
|---|---|---|---|---|
| Starting material amount | 0.5 ~ 5 g | Recovery of low abundant viruses | Long processing time; Difficult in filtration with high mucus samples (such as meconium) | [10, 12–15] |
| | 0.1 ~ 0.3 g | Simple and quick | Lost of low abundant viruses | [17, 28, 29, 31] |
| Suspension buffer | SM buffer | Long-term storage of viruses | | [10, 12–15, 28, 31] |
| | PBS buffer | | | [17, 29] |
| Filtration pore size | 0.20 μm | Better efficiency of removing host and other microbial cells | Lost of viruses larger than 0.20 μm | [13–15, 31] |
| | 0.45 μm | Recovery of viruses larger than 0.20 μm | Less efficiency of removing host and other microbial cells | [12, 29] |
| | 0.45 μm filtration followed by 0.20 μm filtration | | | [10, 17, 28] |
| VLPs enrichment | Centricon Centrifugal Filter | Simple and quick | Proteins from host or other microbial cells cannot be filtered | [13, 31] |
| | CsCl density gradient centrifugation | Better efficiency of removing host and other microbial cells | Long processing time; Limited number of samples that can be processed in parallel | [10, 14, 15] |
| Further purification | Usage of chloroform | Better efficiency of removing host and other microbial cells | Lost of enveloped viruses | [10, 12–15, 17, 28, 31] |
| | No chloroform | Recovery of enveloped viruses | Less efficiency of removing host and other microbial cells | [29] |

must first be reverse transcribed into cDNA, then amplified by sequence-independent, single-primer amplification (SISPA) [35]. or other method [33].

After obtaining sufficient amounts of nucleic acids, virome library construction is similar to standard metagenomic library construction. For example, Illumina Nextera XT Sample Prep kit, which requires only tiny amount of starting materials, is relatively quick, though we note that recovery is not perfectly even—for example, end sequences are typically recovered inefficiently. The Illumina MiSeq and HiSeq platforms are commonly used for virome sequence analysis.

## Wrestling with contamination

Contamination is a challenge when performing metagenomic analysis of samples with low microbial biomass [36, 37]. DNA contamination can come from the laboratory environment, and from commercial reagents. Several studies have characterized the background originating in commercial reagents, and further reported that different kits can bring in different contaminants [36, 37]. Recent studies reported a large number of apparent virus-derived reads from negative control samples in studies of lung bronchoalveolar lavage, serum [33] and feces [31]. In Kim et al. [36], the authors reported numerous reads in a negative control sample which mapped to the phi29 polymerase gene–phi29 polymerase was used to perform GenomiPhi DNA amplification of the samples, suggesting that these reads are likely contamination from the phi29 polymerase protein preparation [36] (i.e. the gene used to manufacture a commercial polymerase came through in the polymerase prep!). Environmental and reagent contamination can be suppressed using ultraclean reagents, but some contamination is probably unavoidable, so it is crucial to use appropriate negative control samples to characterize the background and incorporate results into the interpretation.

## Approaches for analyzing data from virome sequencing studies

Several approaches have been used for analyzing high throughput virome sequence data to identify the composition and types of known viruses and to discover novel viruses. The two approaches involve common steps at the start (Fig. 2). The first step involves removing the adapter sequences which were added during the library preparation stage, using, for example Cutadapt [38]. Next, low quality reads are removed using Trimmomatic [39] or custom scripts. Human reads can then be filtered out using BLAST [40].

Sequence reads can be analyzed individually, or assembled [41–43] into larger "contigs" that represent viral genomes or parts of genomes. The longer contigs provide a longer sequence for similarity searches using BLAST or motifs in inferred protein sequences using Pfam [44, 45]. Use of contigs also allows for more sensitive tracking of viruses over multiple sampling points. Methods for constructing contigs are still being optimized, and multiple

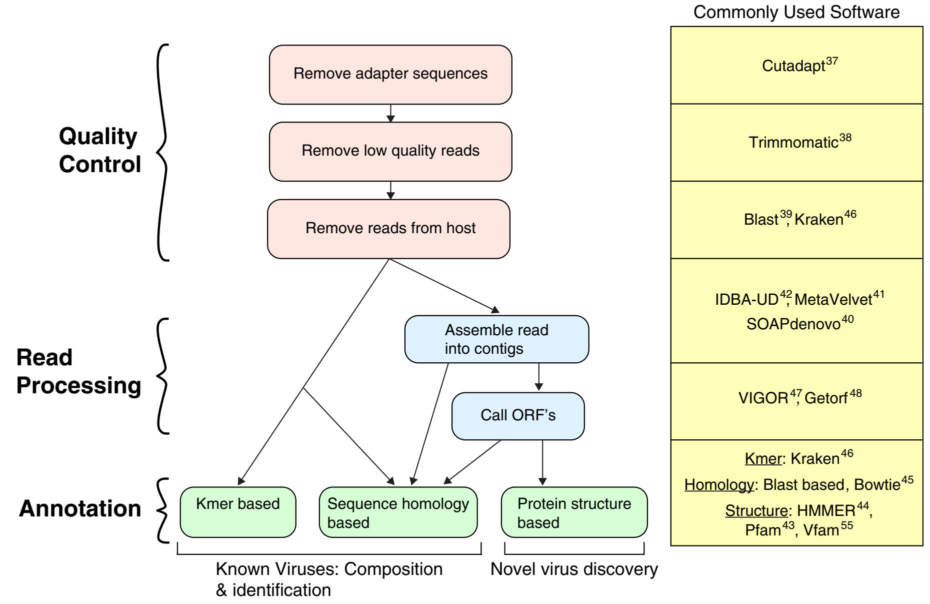

**Fig. 2** Bioinformatic approaches for analyzing the virome sequencing data. The left panel describes the steps in the analysis of the data, while the right panel lists some commonly used tools for performing the corresponding step on the left

challenges remain [46]. For example, sequence heterogeneity and relative abundance of genomes can affect the outcome. Downstream, BLAST [40], Bowtie [47], and Kraken [48] can all be used to detect sequence homology of reads and contigs to reference sequences in the viral database and thus quantify abundance and composition. Open reading frames (ORFs) can also be called [49, 50] on contigs to predict and identify viral genes of interest.

The NCBI Genome database includes the reference whole genome sequences of 7321 viruses. In addition, viral protein sequences are available in Refseq [51], UniProt [52], and custom databases of viral proteins are also available for VLP samples from ocean [53], various geographical habitats [54] or humans [17]. However, alignment to these databases is often challenging when sequence identity is less than 30%. Viruses often accumulate substitutions at high rates [55]–RNA viruses replicate using error prone RNA-dependent RNA polymerases [56], retroviruses use error prone reverse transcriptases [57] and single stranded DNA viruses also show high rates of substitution [55].

These challenges can be addressed by focusing on profile methods for detecting distant homologs of known viral families. The profile methods, specifically those based on hidden markov models (HMM) [45], learn position specific features from sequences and allow for variation at each site under a probabilistic framework. This allows for the query sequence to match the viral family profile HMM if it is evolving like other members in the family, even if it is not highly pairwise similar to any. Here, popular approaches include the protein family database Pfam or virus specific protein family database Vfam [58]. However, Pfam captures only 20% of viral protein families so will not annotate most viral ORFs in a sample. Vfam provides a set of HMMs derived from viral proteins, but does not have detailed annotation of protein function. Thus, further development of these tools would be useful.

Several pipelines [59–64] are available which combine different tools for pre-processing, assembly and annotation. They provide a single step portal for analysis of reads from virome sequencing datasets, using multiple available programs.

None of these tools solve the problem of the viral dark matter (Fig. 1). This is expected given the vast number of viruses in the world and the limited size of available databases. This problem is of less concern for identification and discovery of pathogenic viruses that infect human cells, where there are fewer different types, and these viruses have been closely studied because of their medical importance. However, any study focusing on phage and bacterial dynamics is greatly complicated by the dark matter problem.

## Metagenomic studies of the gut virome

In the sections below, we first review studies that begin to outline the structure of the gut virome and some aspects of its dynamics. Given the interests of the readers of Mobile DNA, we then review two topics on phage mobilization. We first review movement of medically relevant genes between bacterial cells by phage. We focus on a controversy on whether phages are or are not major vehicles for moving antibiotic resistance genes between cells. We then review metagenomic studies documenting movement of whole populations of phages between human individuals during fecal microbial transplantation.

## Composition of the human gut virome

Multiple studies have now investigated the composition of the human gut virome, providing an initial picture of its structure (e. g. [10–17, 28, 31, 65]). As above, researchers have first purified VLPs, then acquired DNA sequence data, allowing assembly and evaluation of contigs. This sketches out aspects of the viral population structure, but a complication is the fact that different viruses are present in distinct abundances. As a result, the most abundant genomes will be sequenced to greater depth, while the rarer genomes will be sparsely covered, or not represented at all. For genomes that are sparsely sequenced, read coverage will be patchy, so that rarer genomes may be represented by multiple contigs, each a fragment of the full genome. Investigators report the number of viral contigs detected, but this is a mixture of full viral genomes and fragments, so the true number of viral variants is challenging to evaluate even roughly. In another approach, the PHAACS program [66] queries how often viral reads assemble together, and uses this to estimate the number of different types. Estimates of human gut populations by PHAACS range from ~2300 to ~8000 phage genotypes. However, implementing this approach requires estimating the mean and variance in genome sizes, which is usually unknown, complicating the analysis.

A simple means of estimating viral abundance is to purify viruses from a weighed amount of stool, then stain with SYBR Gold, which binds nucleic acids, allowing counting of particles. This of course measures all types of viruses as a pool. Such counts are valuable, but we find that RNA virus stain less brightly (unpublished data), and the analysis relies on the premise that all viruses were successfully extracted from a stool sample, both significant limitations. For human stool, counts tend to range from $10^8$ to $10^9$ per gram [67] (our unpublished data); for comparison, the bacterial counts range from $10^{10}$ to $10^{11}$ [68].

Although most viral reads find no attribution of any kind, the minority that do find annotation after alignment to databases allows a provisional accounting of the viral types present. In human stool, the predominant forms are nonenveloped DNA bacteriophages. Tailed phages such as Sipho-, Podo- and Myoviridae are consistently abundant. Microviridae, non-tailed single stranded DNA phages, are also notably abundant, but these are preferentially amplified using MDA (Genomiphi), so that their true abundance in the starting sample is usually unclear without follow up studies.

Assigning VLP contigs to probable microbial hosts is an ongoing challenge. Given a metagenomic sequence sample of viral genomes, say from stool, and a metagenomic analysis of the bacterial taxa present, how do you know who goes with who? Three approaches provide provisional annotation [10, 11, 13–15]. 1) On rare occasions, a VLP contig will closely resemble a database virus with a known host, allowing straightforward attribution. 2) Occasionally a VLP contig will have a reasonably close match to a continuous sequence in a bacterial genome, supporting the idea that the VLP contig corresponds to a temperate phage infecting the queried bacteria. 3) If CRISPR spacers present in a bacterial genome match sequences in a VLP contig from the same environment, it seems reasonable to infer that the virus can infect the CRISPR-containing bacteria. Unfortunately, application of the three methods still usually specifies phage/host relationships for a small minority of VLP contigs in a metagenomic sample. Several groups are developing further methods for use with this problem [69].

Viruses that grow on human cells instead of bacterial cells are typically rare in stool virome samples from healthy subjects. Viral lineages detected include single stranded DNA viruses such as Anelloviruses, Circoviruses, and Parvoviruses, and double-stranded DNA viruses such as Adenoviruses and Papillomaviruses. For RNA viruses in health human stool, viruses of plants seem to predominate, and are inferred to be transients from food. In one memorable study, Pepper mild mottle virus was found to predominate in stool from subjects in California. Extensive detective work showed that the virus was in fact abundant in hot sauce, the apparent source [19].

All these inferences, of course, are greatly complicated by the fact that most genomes in a sample are from viruses that have never been studied. As we become more adept at interrogating the viral dark matter, our thinking on the above points will likely evolve.

## Virome of monozygotic twins and mothers

In one of the earliest comprehensive studies of the human gut virome, Gordon and colleagues [10] investigated the viral component of the human microbiome in healthy individuals using metagenomic sequencing of fecal samples from four pairs of adult female monozygotic twins and their mothers at three time points over a one year period. They found that prophages and temperate phages were abundant in the samples, including Podoviridae, Myoviridae and Siphoviridae families.

They predicted the hosts of the some of the identified VLP contigs using the approaches described above, and found them to be members of the phyla Firmicutes and Bacteroidetes. The majority of the virome was unique to each individual, familial relationships notwithstanding, and showed high inter-personal variability but negligible intra-personal variability over the time period studied. Over 95% of viral genotypes persisted over the one year sampling period [70], and a later study of one healthy adult individual over ~2.5 years showed ~80% persistence [13]. The above studies were ground-breaking, but still the authors could not annotate ~81% of reads, highlighting the importance of the viral dark matter.

### Virome and its response to diet

Gut bacteria are affected by diet [71, 72], so diet is expected to change the composition of phage communities as well. In one study of the dynamics of the human gut virome under a dietary intervention [15], Minot et al. studied fecal samples from six adults on either of two controlled diets for 10 days. Virus-like particles (VLPs) were purified from stool and sequenced, then reads assembled. The authors found each individual harbored a unique and stable virome over the 10 days, suggesting that gut phages are not acquired from food on daily time scales. Individuals on the same diet converged detectably in population composition, suggesting that diet did influence virome composition.

Gordon and colleagues studied [28] the development of the infant virome in healthy and malnourished twins in Malawi. Previous work [73] from the Gordon group had demonstrated that the cellular gut microbiota influences severe acute malnutrition (SAM), so the authors further investigated the role of virome. They sequenced VLPs in fecal samples from 8 pairs of monozygotic and dizygotic twins concordat for healthy growth and 12 twin pairs discordant for SAM over the first three years of life together with their mothers and siblings. The authors developed a machine learning algorithm on virome sequencing reads and identified age discriminatory viruses in healthy twins. They further compared these viruses with those identified from SAM discordant datasets and found phages and eukaryotic viruses belonging to Anelloviridae and Circoviridae families can discriminate discordant from healthy twin pairs. SAM was characterized by a virome community and as well as an immature microbiome. Even the apparently healthy child in the discordant pair had an immature virome, suggesting they may have increased risk for malnutrition. This virome signature was also present after standard therapeutic food therapy for malnutrition, suggesting monitoring the virome may help guide development of improved interventions.

In the sections below we turn to metagenomic studies of phage mobilization. We first review transfer of medically significant gene types between bacteria, then movement of whole viral communities between human individuals during fecal microbial transplantation.

### Transport and integration of medically important genes by phage

Temperate bacteriophage can transport genes between bacteria and install them in the bacterial genome by integration [74, 75]. These genes are then inherited like normal bacterial genes during DNA replication and cell division. Upon sensing of a suitable inducing signal such as DNA damage, the prophage can excise, replicate lytically, and release progeny capable of infecting new cells [76–81]. Thus, cells harboring prophages— "lysogens"—can show novel phenotypic characteristics resulting from expression of genes on prophages, some of which are medically relevant.

For example, phages are well known to transport toxin genes between bacterial cells [82–84]. Shiga toxin, cholera toxin, and numerous others are carried on temperate phage, so that transduction renders lysogenic bacteria toxin producers. Integration of the phage genome into the bacterial genome can take place via either phage-encoded integrases (shiga toxin) [84] or by hijacking host cell recombination machinery (cholera toxin) [83]. Virome studies are just beginning to report the global frequency of occurrence of such toxin genes in different environments [82]. Other gene types are also known to influence human health [25].

Less clear has been the extent to which antibiotic resistance genes have been transferred between bacteria via phage. Historically, phage transduction has been viewed as only a minor contributor to transmission of antibiotic resistance genes, with transformation and particularly conjugation mediating transfer to a far greater extent [75]. However, a recent metagenomic study suggested that phage commonly encode antibiotic resistance genes, and that in mice the frequency of antibiotic resistance genes on phage actually increases with antibiotic treatment [85]. This supported a disturbing model in which antibiotic treatment actually caused wholesale mobilization of resistance genes via phage.

However, a recent reanalysis of annotation methods suggested a technical explanation. If thresholds for annotating antibiotic resistance genes are excessively permissive, then many calls may be erroneous misattribution of genes with other functions. Enault et al. [86] carried out a careful comparison of annotation thresholds for calling antibiotic resistance genes, combined with functional tests, and suggested that in fact the thresholds used by Modi et al. were far too permissive, so that far fewer resistance genes were present than initially thought. Analysis of fully sequenced phage genomes yielded only four clear examples of well-supported antibiotic resistance genes [86]. More data in this area would be helpful, but it now seems that the original picture may have been correct, and phage are only rare carriers of antibiotic resistance genes.

It is also rare to find transposons integrated into phage genomes. Thus, a major piece of the apparatus important for transmissible antibiotic resistance is again rare in phage. Possibly this is due to packaging efficiency: viral capsids can incorporate only a certain amount of nucleic acid, and lengthening viral genomes by transposon insertion may result in genomes that are incorporated relatively inefficiently.

### Movement of phage between humans during fecal microbial transplantation

Fecal microbiota transplant has been successful in treatment of relapsing *Clostridium difficile* (*C. difficile*) infections [87]. FMT treatment appears to work by restoring a more normal anaerobic gut community, though measurements typically show that the new communities in patients are complex mixtures of strains from donor, recipient, and new acquisition [88]. The general behavior and possible contribution of the virome in FMT is just starting to be investigated.

Chehoud et al. [31] sequenced the virome from a case series in which feces from a single donor was used to treat three children with ulcerative colitis (UC). Recipients received multiple FMT treatments over a 6 to 12 weeks' period. Possible transient clinical benefit was observed [89]. The authors sequenced donor and recipient VLP samples, and assembled contigs from the reads. Multiple donor viral contigs were detected in the donor and in each recipient. Up to 42 donor contigs were detected in recipients, some annotating to specific bacteriophage families, documenting extensive transfer of phage communities. Chehoud et al. also investigated features associated with preferential transfer of viruses from donors to recipients, and found signatures of lysogeny in the transmitted viruses–the two most frequently transferred gene types were associated with temperate phage replication, and *Siphoviridae*, the group including lambda, were transferred with high efficiency. This led to the proposal that lysogeny may exist in part to assist in phage dispersal between environments.

More recently, Zuo and colleagues [65] investigated the role of the virome in FMT treatment for *C. difficile* infection. They sequenced the virome from 24 subjects with *C. difficile*, of whom 9 were treated with FMT and 5 received standard care with antibiotics, and 20 healthy controls. They found that before treatment patients with *C. difficile* had a higher abundance of phages from *Caudiovirales* (tailed bacteriophages) but lower diversity, richness and evenness compared to healthy controls. Following FMT treatment, subjects who responded showed an increased abundance of *Caudiovirales* contigs from the donor compared with those who did not respond. This raises the intriguing possibility that phage may be involved in successful FMT, possibly consistent with a published pilot study in which fecal extracts lacking bacteria were potentially effective in treating Clostridium difficile infection [90].

## Conclusions

Recognition of the vast phage populations associated with humans prompts numerous questions on their biology. How many different kinds are there? What are their replication styles and rates? How do genes transported by phage influence bacterial phenotypes relevant to human health? Most broadly, how do phage affect human welfare?

We are starting to see proposals for associations between large groups of phages and specific human disease. For example, Caudovirales have been associated with human inflammatory bowel disease in some [17] but not all [91] studies. The Caudovirales are a large and heterogenous order—it seems surprising that they should be behaving similarly as a group, but mechanisms have been proposed to explain this [17]. Similarly, as mentioned above, Caudovirales abundance has been associated with success in fecal microbial transplantation [65], another intriguing idea that awaits confirmation in further data sets.

Phage-mediated DNA mobilization no doubt also strongly influences human-associated communities and thereby human health. Phage were recently shown to move DNA between gut Salmonella strains in mice in response to induction by reactive oxygen species [92]. Likely myriad phage in gut move between bacterial species in response to further inducing agents characteristic of the gut environment, many of which are likely to be unidentified so far. It will be valuable to characterize transfer in more detail in human-associated settings. Finally, movement of whole phage populations between individuals are just starting to be studied, with initial focus on FMT due to the experimental accessibility.

Recent work provides a new window on an old problem, which is the role of lysogeny in phage ecology [93]. Rohwer and colleagues have suggested [1] a "Piggyback-the-Winner" model, where lysogeny is favored at high microbial density. This is in contrast with the earlier "Kill-the-Winner" model [94, 95], which suggests that once a microbial host achieves a high density, it is increasingly preferentially targeted by a predator phage which replicates on the predominant strain. The abundant strain then decreases in relative proportion, resulting in increased microbial diversity of the prey community, thus emphasizing the importance of lytic growth. Piggyback-the-winner suggests that phage actually replicate more efficiently in many environments as a prophage installed in successful bacteria. Recent studies [93, 96] have also highlighted the role of lysogeny in mediating resistance to phage superinfections via phage-encoded phage resistance genes encoded on prophage. In addition, as mentioned above, studies of FMT suggest that lysogeny may also assist in phage dispersal. Thus, contemporary virome studies lead us to think about the role of lysogeny in several new ways.

We end with a conjecture on the nature of the viral dark matter [16]. Why is such a large fraction of phage DNA sequence unlike any previously studied? One idea is that genomes of DNA phage are under pressure to change their primary sequences in response to pressure from restriction endonucleases and CRISPR systems. Ongoing host-virus competition, played out at a replication rate as fast as 20 min per cycle, will drive high rates of sequence diversification. If this is then multiplied over the estimated $10^{31}$ viral particles on Earth, it becomes easier to understand how phage have diversified to an extreme degree. A corollary is that despite the rapid drift in the primary DNA sequence, protein structure and function may be more conserved. In a few cases there are multiple X-ray structures for different phage proteins that carry out conserved functions, allowing assessment of their resemblance. For the phage repressor and Cro proteins, which are important in regulating lysogeny, the DNA sequences from lambda, 434 and P22 have little resemblance (median identity 34%), and even less resemblance at the protein level (median identity 17%) [97]. However, the encoded proteins show generally similar structures, dominated by the helix-turn-helix DNA binding motif and supporting alpha-helical secondary structures [98–102]. If this is generalizable, then perhaps once phage protein structures and functions are better worked out, understanding the viral dark matter will become less daunting.

### Abbreviations

*C. difficile*: Clostridium Difficile; CRISPR: Clustered interspersed short palindromic repeats; CsCl: Cesium chloride; EM: Electron Microscopy; FMT: Fecal Microbiota Transplant; SAM: Severe acute malnutrition; VLP: Virus-like particle

### Acknowledgements

We thank Laurie Zimmerman for assistance with the figures.

### Funding

This work was supported by AI082020-05A1 (FDB), HL117798–05 (Fitzgerald), HL113252–04 (Collman, Bushman, Christie), the Penn Center for AIDS Research P30 AI 045008, Pennsylvania Department of Health SAP 4100068710, and the PennCHOP Microbiome Program. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data, or in writing the manuscript.

### Availability of data and materials

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

### Authors' contributions

All authors contributed to conceiving and writing this review article. All authors have read the manuscript and consented to publication. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

[1]Department of Microbiology, University of Pennsylvania School of Medicine, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA. [2]Division of Gastroenterology, Hepatology, and Nutrition, Children's Hospital of Philadelphia, Philadelphia, PA 19104-4319, USA.

Received: 25 July 2017 Accepted: 24 September 2017
Published online: 03 October 2017

### References

1. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobian-Guemes AG, et al. Lytic to temperate switching of viral communities. Nature. 2016;531:466–70.
2. Proctor LM. Advances in the study of marine viruses. Microsc Res Tech. 1997;37:136–61.
3. Proctor LM, Okubo A, Fuhrman JA. Calibrating estimates of phage-induced mortality in marine bacteria: Ultrastructural studies of marine bacteriophage development from one-step growth experiments. Microb Ecol. 1993;25:161–82.
4. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449:804–10.
5. Wu GD, Lewis JD, Hoffmann C, Chen YY, Knight R, Bittinger K, et al. Sampling and pyrosequencing methods for characterizing bacterial communities in the human gut using 16S sequence tags. BMC Microbiol. 2010;10:206.
6. Human Microbiome Project C. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
7. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. Science. 2015; 348:1261498.
8. Hoffmann C, Dollive S, Grunberg S, Chen J, Li H, Wu GD, et al. Archaea and fungi of the human gut microbiome: correlations with diet and bacterial residents. PLoS One. 2013;8:e66019.
9. Zaneveld J, Turnbaugh PJ, Lozupone C, Ley RE, Hamady M, Gordon JI, et al. Host-bacterial coevolution and the search for new drug targets. Curr Opin Chem Biol. 2008;12:109–14.
10. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466:334–8.
11. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol. 2012;10:607–17.
12. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proc Natl Acad Sci U S A. 2013;110:20236–41.
13. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. Rapid evolution of the human gut virome. Proc Natl Acad Sci U S A. 2013;110: 12450–5.
14. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD. Hypervariable loci in the human gut virome. Proc Natl Acad Sci U S A. 2012;109:3962–6.
15. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 2011;21:1616–25.
16. Minot S, Wu GD, Lewis JD, Bushman FD. Conservation of gene cassettes among diverse viruses of the human gut. PLoS One. 2012;7:e42342.
17. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160:447–60.

18. Young JC, Chehoud C, Bittinger K, Bailey A, Diamond JM, Cantu E, et al. Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. Am J Transplant Off J Am Soc Transplant Am Soc Transplant Surg. 2015;15:200–9.
19. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. PLoS Biol. 2006;4:e3.
20. Rohwer F, Segall AM. In retrospect: A century of phage lessons. Nature. 2015;528:46–8.
21. Cadwell K. The virome in host health and disease. Immunity. 2015;42:805–13.
22. Viertel TM, Ritter K, Horz HP. Viruses versus bacteria-novel approaches to phage therapy as a tool against multidrug-resistant pathogens. J Antimicrob Chemother. 2014;69:2326–36.
23. Virgin HW, Wherry EJ, Ahmed R. Redefining chronic viral infection. Cell. 2009;138:30–50.
24. Sauvage V, Eloit M. Viral metagenomics and blood safety. Transfus Clin Biol. 2016;23:28–38.
25. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003;113:171–82.
26. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
27. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc. 2009;4:470–83.
28. Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, et al. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A. 2015;112:11941–6.
29. Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, et al. Early life dynamics of the human gut virome and bacterial microbiome in infants. Nat Med. 2015;21:1228–34.
30. Kleiner M, Hooper LV, Duerkop BA. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics. 2015;16:7.
31. Chehoud C, Dryga A, Hwang Y, Nagy-Szakal D, Hollister EB, Luna RA, et al. Transfer of Viral Communities between Human Individuals during Fecal Microbiota Transplantation. MBio. 2016;7:e00322.
32. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540:539–43.
33. Abbas AA, Diamond JM, Chehoud C, Chang B, Kotzin JJ, Young JC, et al. The Perioperative Lung Transplant Virome: Torque Teno Viruses Are Elevated in Donor Lungs and Show Divergent Dynamics in Primary Graft Dysfunction. Am J Transplant. 2017;17:1313–24.
34. Kim KH, Chang HW, Nam YD, Roh SW, Kim MS, Sung Y, et al. Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol. 2008;74:5975–85.
35. Wang D, Urisman A, Liu Y-T, Springer M, Ksiazek TG, Erdman DD, et al. Viral Discovery and Sequence Recovery Using DNA Microarrays. PLoS Biol. 2003;1:e2.
36. Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, et al. Optimizing methods and dodging pitfalls in microbiome research. Microbiome. 2017;5:52.
37. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
38. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10.
39. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014;30:2114–20.
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
41. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18.
42. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40:e155.
43. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
44. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290–301.
45. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29–37.
46. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14:157–67.
47. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
48. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
49. Wang S, Sundaram JP, Spiro D. VIGOR, an annotation program for small viral genomes. BMC Bioinformatics. 2010;11:451.
50. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics TIG. 2000;16:276–7.
51. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–D45.
52. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. Nucleic Acids Res. 2011;39:D576–82.
53. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537:689–93.
54. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536:425–30.
55. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9:267–76.
56. Lauring AS, Frydman J, Andino R. The role of mutational robustness in RNA virus evolution. Nat Rev Microbiol. 2013;11:327–36.
57. Svarovskaia ES, Cheslock SR, Zhang W-H, Hu W-S, Pathak VK. Retroviral mutation rates and reverse transcriptase fidelity. Front Biosci J Virtual Libr. 2003;8:d117–34.
58. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL, Chase-Topping M. Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. PLoS One. 2014;9:e105067.
59. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, et al. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology. 2017;503:21–30.
60. Roux S, Tournayre J, Mahul A, Debroas D, Enault F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics. 2014;15:76.
61. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci. 2012;6:427–39.
62. Rampelli S, Soverini M, Turroni S, Quercia S, Biagi E, Brigidi P, et al. ViromeScan: a new tool for metagenomic viral community profiling. BMC Genomics. 2016;17:165.
63. Lorenzi HA, Hoover J, Inman J, Safford T, Murphy S, Kagan L, et al. TheViral MetaGenome Annotation Pipeline (VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. Stand Genomic Sci. 2011;4:418–29.
64. Ho T, Tzanetakis IE. Development of a virus detection and discovery pipeline using next generation sequencing. Virology. 2014;471-473:54–60.
65. Zuo T, Wong SH, Lam K, Lui R, Cheung K, Tang W et al. Bacteriophage transfer during faecal microbiota transplantation in Clostridium difficile infection is associated with treatment outcome. Gut. 2017;0:1–10.
66. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, et al. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC bioinformatics. 2005;6:41.
67. Kim M-S, Park E-J, Roh SW, Bae J-W. Diversity and abundance of single-stranded DNA viruses in human feces. Appl Environ Microbiol. 2011;77:8062–70.
68. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14:e1002533.
69. Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun. 2017;8:15955.
70. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
71. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. Sci Transl Med. 2009;1:6ra14.
72. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2013;505:559–63.

73. Smith MI, Yatsunenko T, Manary MJ, Trehan I, Mkakosya R, Cheng J, et al. Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor. Science. 2013;339:548–54.
74. Ptashne M, Jeffrey A, Johnson AD, Maurer R, Meyer BJ, Pabo CO, et al. How the lambda repressor and cro work. Cell. 1980;19:1–11.
75. Bushman F: Lateral DNA transfer : mechanisms and consequences. New York: Cold Spring Harbor Laboratory Press; 2002.
76. Guarente L, Nye JS, Hochschild A, Ptashne M. Mutant lambda phage repressor with a specific defect in its positive control function. Proc Natl Acad Sci U S A. 1982;79:2236–9.
77. Hochschild A, Irwin N, Ptashne M. Repressor structure and the mechanism of positive control. Cell. 1983;32:319–25.
78. Bushman FD, Ptashne M. Activation of transcription by the bacteriophage 434 repressor. Proc Natl Acad Sci U S A. 1986;83:9353–7.
79. Bushman FD, Ptashne M. Turning lambda Cro into a transcriptional activator. Cell. 1988;54:191–7.
80. Bushman FD, Shang C, Ptashne M. A single glutamic acid residue plays a key role in the transcriptional activation function of lambda repressor. Cell. 1989;58:1163–71.
81. Ptashne M: A genetic switch : phage lambda revisited. New York: Cold Spring Harbor Laboratory Press; 2004.
82. Casas V, Maloy S. Role of bacteriophage-encoded exotoxins in the evolution of bacterial pathogens. Future Microbiol. 2011;6:1461–73.
83. McLeod SM, Kimsey HH, Davis BM, Waldor MK. CTXphi and Vibrio cholerae: exploring a newly recognized type of phage-host cell relationship. Mol Microbiol. 2005;57:347–56.
84. Herold S, Karch H, Schmidt H. Shiga toxin-encoding bacteriophages–genomes in motion. Int J Med Microbiol. 2004;294:115–21.
85. Modi SR, Lee HH, Spina CS, Collins JJ. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. Nature. 2013;499:219–22.
86. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. ISME J. 2017;11:237–47.
87. Rohlke F, Stollman N. Fecal microbiota transplantation in relapsing Clostridium difficile infection. Ther Adv Gastroenterol. 2012;5:403–20.
88. Li SS, Zhu A, Benes V, Costea PI, Hercog R, Hildebrand F, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science. 2016;352:586–9.
89. Kellermayer R, Nagy-Szakal D, Harris RA, Luna RA, Pitashny M, Schady D, et al. Serial fecal microbiota transplantation alters mucosal gene expression in pediatric ulcerative colitis. Am J Gastroenterol. 2015;110:604–6.
90. Ott SJ, Waetzig GH, Rehman A, Moltzau-Anderson J, Bharti R, Grasis JA, et al. Efficacy of Sterile Fecal Filtrate Transfer for Treating Patients With Clostridium difficile Infection. Gastroenterology. 2017;152:799–811. e7
91. Lewis JD, Chen EZ, Baldassano RN, Otley AR, Griffiths AM, Lee D, et al. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. Cell Host Microbe. 2015;18:489–500.
92. Diard M, Bakkeren E, Cornuault JK, Moor K, Hausmann A, Sellin ME, et al. Inflammation boosts bacteriophage transfer between Salmonella spp. Science. 2017;355:1211–5.
93. Bondy-Denomy J, Qian J, Westra ER, Buckling A, Guttman DS, Davidson AR, et al. Prophages mediate defense against phage infection through diverse mechanisms. ISME J. 2016;10:2854–66.
94. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, et al. Viral and microbial community dynamics in four aquatic environments. ISME J. 2010;4:739–51.
95. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, et al. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7:828–36.
96. Dedrick RM, Jacobs-Sera D, Bustamante CA, Garlena RA, Mavrich TN, Pope WH, et al. Prophage-mediated defence against viral attack and viral counter-defence. Nat Microbiol. 2017;2:16251.
97. Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO. Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature. 1982;298:447–51.
98. Ohlendorf DH, Anderson WF, Lewis M, Pabo CO, Matthews BW. Comparison of the structures of cro and lambda repressor proteins from bacteriophage lambda. J Mol Biol. 1983;169:757–69.
99. Anderson JE, Ptashne M, Harrison SC. A phage repressor-operator complex at 7 A resolution. Nature. 1985;316:596–601.
100. Bushman FD, Anderson JE, Harrison SC, Ptashne M. Ethylation interference and X-ray crystallography identify similar interactions between 434 repressor and operator. Nature. 1985;316:651–3.
101. Wolberger C, Dong YC, Ptashne M, Harrison SC. Structure of a phage 434 Cro/DNA complex. Nature. 1988;335:789–95.
102. Sevilla-Sierra P, Otting G, Wuthrich K. Determination of the nuclear magnetic resonance structure of the DNA-binding domain of the P22 c2 repressor (1 to 76) in solution and comparison with the DNA-binding domain of the 434 repressor. J Mol Biol. 1994;235:1003–20.