AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

COMMENTARY

Phage Diversity in the Human Gut Microbiome: a Taxonomist's Perspective

Evelien M. Adriaenssens[a]

[a]Quadram Institute Bioscience, Norwich, United Kingdom

ABSTRACT Bacteriophages (phages) have been known for over a century, but only in the last 2 decades have we really come to appreciate how abundant and diverse they are. With that realization, research groups across the globe have shown the importance of phage-based processes in a myriad of environments, including the global oceans and soils, and as part of the human microbiome. Through advances in sequencing technology, genomics, and bioinformatics, we know that the morphological diversity of bacteriophages originally used for taxonomy is eclipsed by their genomic diversity. Because we currently do not have a complete taxonomic framework or naming scheme to describe this diversity, crucial information from virome and microbiome studies is being lost. In this commentary, I will discuss recent advances in taxonomy and its importance for studies of the microbiome with examples of the human gut phageome and make recommendations for future analyses.

KEYWORDS human gut virome, microbiome, phage taxonomy, phageome, virome

RECENT CHANGES IN PHAGE TAXONOMY AND THEIR IMPLICATIONS FOR MICROBIOME ANALYSES

In 2020, a major step was taken in virus taxonomy with the implementation of higher ranks, the so-called megataxonomy of viruses (1, 2), providing 15 hierarchical ranks in which to classify all viruses. The known diversity of phages is now spread over four realms (*Duplodnaviria*, *Monodnaviria*, *Varidnaviria*, and *Riboviria*) that encompass six kingdoms and seven phyla (Fig. 1). The most commonly isolated phages, double-stranded DNA (dsDNA) tailed bacteriophages with a HK97-like major capsid protein, are unified in the class *Caudoviricetes*, at the time of writing equivalent with the order *Caudovirales*. At the family level, which is often used as a bin to visualize metagenomics data, the phage taxonomy is undergoing a rapid revolution from morphology-based classification in favor of a genome-based classification (3). As a result, many new families are being created so that members of the same family share a set of core genes, which is not the case with the classification into the families *Myoviridae*, *Podoviridae*, and *Siphoviridae*, which are scheduled to be abolished. At the ranks of species and genus, nucleotide identity-based demarcation criteria have been implemented that allow for systematic binning of metagenome data at these ranks (3–6). These levels are the most well-curated and comprehensive, reflected in the high number of proposals describing new genera in recent years (7–9). However, since phage (and all virus) taxonomy is done *post hoc*, i.e., new phage isolates are described and published first and only then classified by committee, the latest taxonomy database will always lag behind the "known" phage diversity.

As a result of the changes to and limitations of taxonomy, the current phage taxonomy database, as described on the website of the ICTV (International Committee on Taxonomy of Viruses) (ictvonline.org) and implemented by NCBI Taxonomy (10), is what I can only describe as a bit of a mixed bag. Given the large amounts of manual curation involved with classification and nomenclature, some parts of the phage sequence space have been tackled recently and are thus clearly defined, while others

Citation Adriaenssens EM. 2021. Phage diversity in the human gut microbiome: a taxonomist's perspective. mSystems 6:e00799-21. https://doi.org/10.1128/mSystems.00799-21.

Address correspondence to evelien.adriaenssens@quadram.ac.uk.

Conflict of Interest Disclosures: E.M.A. reports grants from the Biotechnology and Biological Sciences Research Council (BBSRC) during the conduct of the study. She is the Chair of the Bacterial Viruses Subcommittee of the International Committee on Taxonomy of Viruses (ICTV).

The views expressed in this article do not necessarily reflect the views of the journal or of ASM.

This article is part of a special series sponsored by Floré.

Published 17 August 2021

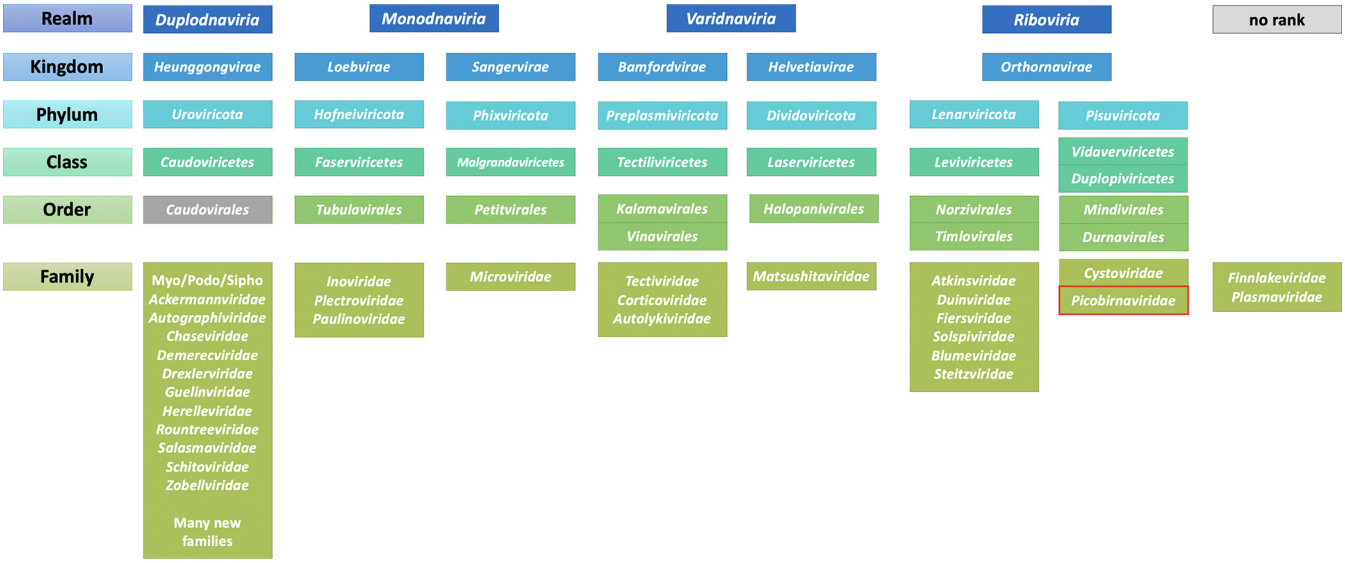

FIG 1 Overview of the virus ranks containing bacteriophages as of Master Species List 36 (https://talk.ictvonline.org/files/master-species-lists/m/msl/12314 [accessed June 2021]). The order *Caudovirales* is indicated in gray as it is scheduled for deletion. The family *Picobirnaviridae* outlined in red was originally recognized as a family of animal-associated viruses, but is now bioinformatically predicted to be made up of bacteriophages.

are not. This poses a lot of issues for the correct interpretation of microbiome/virome data, exactly because family-level descriptions are so often used (including in the past by myself). Unfortunately, some of these family-level analyses are wrong—for now—and should be avoided or at the very least manually curated, which I will explore in the example below.

AN EXAMPLE OF A HEALTHY HUMAN GUT PHAGEOME: WHERE CAN THE ANALYSES GO WRONG?

In this example, I am using three distinct phage communities extracted from metagenome sequencing data sets from fecal samples from three healthy individuals (data derived from T. Brown and E.M. Adriaenssens, unpublished data). For each sample/individual, we assembled and validated the phage genomes using megahit and VirSorter, respectively (11, 12). Figure 2 shows two different analyses and visualizations of the same data: (i) heatmap of a reference-based assignment of contigs to a viral family using Diamond and Megan (13, 14), (ii) network representation of contigs and reference genomes as nodes (circles) connected by edges that represent shared protein clusters using vConTACT2 and the INPHARED pipeline (15, 16). What is immediately obvious from this comparison is that the dsDNA families *Myoviridae*, *Podoviridae*, and *Siphoviridae* and also the single-stranded DNA (ssDNA) family *Microviridae*, which are in a single bin in the heatmap (Fig. 2A) do not represent the phage sequence space well as they are separated across multiple clusters in the network (Fig. 2B). With the current taxonomic organization, two phages can belong to the same family and share no core proteins (or at least none that we can detect with sequence-based tools). The newer genome-based families are more cohesive across the network, but not the family *Autographiviridae*, which may get split further.

There are additional interesting observations that can be derived from this example. While the phage communities in the three healthy individuals are similar, they are not identical. There are also multiple clusters of related phages that bear no resemblance with database phages, which in the analysis in Fig. 2A are all classed together in the "Not assigned" bin, losing resolution. Where the two analyses are in agreement is the observation that siphoviruses dominate the gut phageome.

Given the realities of phage taxonomy and microbiome analyses, I can make the following recommendations.

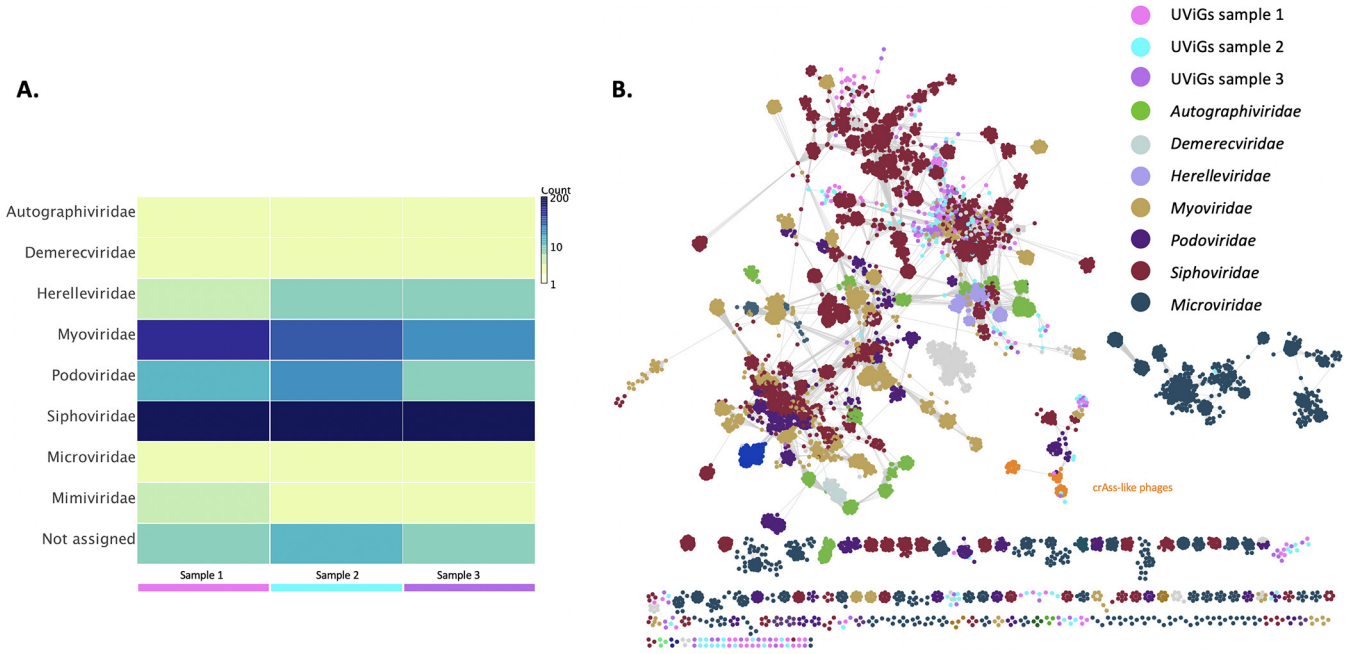

FIG 2 Phage diversity analysis of three gut phageome samples. (A) Heatmap of the number of uncultivated virus genomes (UViGs) per family grouped at the viral family level using Diamond BLASTX against the viral RefSeq database release 99 and lowest common ancestor assignment using Megan6. (B) vConTACT2 network analysis of UViGs from three samples and all published complete phage genomes of the INPHARED pipeline on 24 January 2021. Families are colored according to the INPHARED metadata, with selected families indicated in the legend. Archaeal viruses and the taxa of nontailed phages except for microviruses were removed.

- Do not rely on automated family-level binning approaches for analysis of the phageome
- Use different clustering methods for diversity analyses:
  - Nucleotide level clustering for species (95%) and genus (70%)
  - Shared predicted protein content for subfamilies and families
  - Deep rooted phylogenies of marker genes such as the terminase large subunit (*Caudoviricetes*) or capsid proteins for higher-order classifications
- Use multiple tools for exploration to reduce biases
- Take advantage of additional databases for higher resolution analyses (see examples below)

GLOBAL HUMAN GUT PHAGE DIVERSITY AND crAssphages: ARE WE SPEAKING THE SAME LANGUAGE?

The recent surge in interest in bacteriophage research has led to the creation of a number of overlapping or competing databases describing the gut virome, which allow for additional resolution of gut phageome analyses. The Gut Virome Database (GVD) contains 33,242 viral populations from 1,986 individuals (17). The Cenote Human Virome Database (CHVD) comprises 45,033 viral operational taxonomic units (vOTUs) from all human body sites (18). The human Gut Phage Database is currently the largest gut-specific database with 142,809 nonredundant phage genomes assembled from 28,060 metagenomes and 2,898 bacterial genomes, of which 13,429 were classified as complete and a further 27,999 were classified as high quality by CheckV (19, 20). Another recent study assembled 3,738 complete phage genomes from 5,742 metagenomes (21). The most recent database is the Metagenomic Gut Virus (MGV) catalogue

containing 189,680 (partial) genomes grouped into 54,118 species-level vOTUs (22). The MGV paper recognizes overlap and complementarity of the different databases and highlights the need for a unified and standardized resource, a sentiment I echo with enthusiasm.

In the papers describing these databases, often specific clades of phages are highlighted. For example, the GVD describes 70 crAssphage populations clustered into 12 viral clusters, but no single population shared across individuals (17). This is echoed by the analyses of the CHVD and GPD (18, 19), with the latter identifying a new clade dubbed Gubaphage that is distantly related to crAss-like phages. These descriptions across multiple publications and databases leave us in a Babel-like situation that, for instance, leaves us pondering what the term "crAssphage" actually means. When first described, it was posited as the most abundant human gut-associated phage (23, 24). However, the first cultured crAssphage, *Bacteriodes* phage phicrAss001 showed no nucleotide sequence similarity with the original crAssphage (25). Combining information from metagenomics studies and culturing approaches and driven by a collaboration across multiple research groups, the newly formed "Crassvirales Study Group" of the ICTV has submitted a proposal to create a new order, called *Crassvirales*, divided into multiple families, genera, and species (2021.022B.v1.Crassvirales, https://talk.ictvonline.org/files/proposals/taxonomy_proposals_prokaryote1/ [accessed June 2021]), allowing a taxonomic framework to facilitate the semantics associated with this group of phages (indicated in orange in Fig. 2B). It is my hope that this classification will normalize descriptions of crAss-like phages across publications and facilitate our understanding of this highly interesting group of phages.

CONCLUSIONS AND PERSPECTIVES

In my—perhaps biased—opinion, both the phage community and the microbiome community need a well-curated genome-based taxonomic classification framework for phages. Put more strongly, taxonomy is the language that binds us together and will allow us to understand each other's studies. In future, it is my hope that we can use the taxonomic framework to identify multiple sets of phages that are of importance to human health and disease, whether they are biomarkers for a healthy gut, indicative of a diseased state, or candidates for phage therapy. While this analysis was focused on the human gut, the taxonomic framework is not and will be essential in any environment.

I will leave the reader with three questions that we, as a community, need to answer so that we can understand each other across diverging fields of phage-related research:

- What is a phage?
- What is a viral family?
- When can we confidently say that a phage (or other type of virus) is present in a sample?

ACKNOWLEDGMENTS

I thank Betty Kutter, Teagan Brown, Claire Elek, and Andrea Telatin for their assistance with the human phage study experiments, and the ICTV Bacterial Viruses Subcommittee members and Crassvirales Study Group members for fruitful collaborations. The human samples were processed under the QIB Colon Model ethics, HRGC reference IFR01/2015.

I gratefully acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Program Gut Microbes and Health BB/R012490/1 and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356, and a BBSRC Flexible Talent Mobility Grant BB/R506552/1 to the Quadram Institute Bioscience.

E.M.A. is the Chair of the Bacterial Viruses Subcommittee of the International Committee on Taxonomy of Viruses (ICTV). The funders had no role in the design of the study, its analyses or the decision to publish.

REFERENCES

1. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, Zerbini FM, Kuhn JH. 2020. Global organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev 84:e00061-19.
2. International Committee on Taxonomy of Viruses Executive Committee. 2020. The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. Nat Microbiol 5:668–674. https://doi.org/10.1038/s41564-020-0709-x.
3. Turner D, Kropinski AM, Adriaenssens EM. 2021. A roadmap for genome-based phage taxonomy. Viruses 13:506. https://doi.org/10.3390/v13030506.
4. Adriaenssens E, Brister JR. 2017. How to name and classify your phage: an informal guide. Viruses 9:70. https://doi.org/10.3390/v9040070.
5. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonté JM, Lee K-B, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Páez-Espino D, Petit M-A, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, et al. 2019. Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol 37:29–37. https://doi.org/10.1038/nbt.4306.
6. Moraru C, Varsani A, Kropinski AM. 2020. VIRIDIC—a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses 12:1268. https://doi.org/10.3390/v12111268.
7. Adriaenssens EM, Sullivan MB, Knezevic P, van Zyl LJ, Sarkar BL, Dutilh BE, Alfenas-Zerbini P, Łobocka M, Tong Y, Brister JR, Moreno Switt AI, Klumpp J, Aziz RK, Barylski J, Uchiyama J, Edwards RA, Kropinski AM, Petty NK, Clokie MRJ, Kushkina AI, Morozova VV, Duffy S, Gillis A, Rumnieks J, Kurtböke İ, Chanishvili N, Goodridge L, Wittmann J, Lavigne R, Jang HB, Prangishvili D, Enault F, Turner D, Poranen MM, Oksanen HM, Krupovic M. 2020. Taxonomy of prokaryotic viruses: 2018–2019 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol 165:1253−1260. https://doi.org/10.1007/s00705-020-04577-8.
8. Adriaenssens EM, Wittmann J, Kuhn JH, Turner D, Sullivan MB, Dutilh BE, Bin JH, van Zyl LJ, Klumpp J, Lobocka 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. https://doi.org/10.1007/s00705-018-3723-z.
9. Adriaenssens EM, Krupovic M, Knezevic P, Ackermann H-W, Barylski J, Brister JR, Clokie MRC, Duffy S, Dutilh BE, Edwards RA, Enault F, Bin JH, Klumpp J, Kropinski AM, Lavigne R, Poranen MM, Prangishvili D, Rumnieks J, Sullivan MB, Wittmann J, Oksanen HM, Gillis A, Kuhn JH. 2017. Taxonomy of prokaryotic viruses: 2016 update from the ICTV Bacterial and Archaeal Viruses Subcommittee. Arch Virol 162:1153–1157. https://doi.org/10.1007/s00705-016-3173-4.
10. Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I. 2020. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford) 2020: baaa062. https://doi.org/10.1093/database/baaa062.
11. Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
12. Roux S, Enault F, Hurwitz BL, Sullivan MB. 2015. VirSorter: mining viral signal from microbial genomic data. PeerJ 3:e985. https://doi.org/10.7717/peerj.985.
13. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
14. Huson DH, Weber N. 2013. Microbial community analysis using MEGAN. Methods Enzymol 531:465–485. https://doi.org/10.1016/B978-0-12-407863-5.00021-6.
15. Bin JH, 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.
16. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie MR, Stekel DJ, Hobman J, Jones MA, Millard AD. 1 May 2021. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of phage genomes. bioRxiv https://doi.org/10.1101/2021.05.01.442102.
17. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. 2020. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell Host Microbe 28:724–740.e8. https://doi.org/10.1016/j.chom.2020.08.003.
18. Tisza MJ, Buck CB. 2021. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. Proc Natl Acad Sci U S A 118:e2023202118. https://doi.org/10.1073/pnas.2023202118.
19. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.
20. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides N. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578–585. https://doi.org/10.1038/s41587-020-00774-7.
21. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, Pevzner P, Koonin EV. 2021. Thousands of previously unknown phages discovered in whole-community human gut metagenomes. Microbiome 9:78. https://doi.org/10.1186/s40168-021-01017-w.
22. Nayfach S, Páez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, Proal AD, Fischbach MA, Bhatt AS, Hugenholtz P, Kyrpides NC. 2021. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol 6:960−970. https://doi.org/10.1038/s41564-021-00928-6.
23. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, Barr JJ, Speth DR, Seguritan V, Aziz RK, Felts B, Dinsdale EA, Mokili JL, Edwards RA. 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun 5:4498. https://doi.org/10.1038/ncomms5498.
24. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, Draper LA, Gonzalez-Tortuero E, Ross RP, Hill C. 2018. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe 24:653–664.e6. https://doi.org/10.1016/j.chom.2018.10.002.
25. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, Hill C. 2018. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. Nat Commun 9:4781. https://doi.org/10.1038/s41467-018-07225-7.