CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Ten quick tips for classifying unknown bacteriophage
Authors	Bastiaanssen, Fabian T.S.;Hill, Colin;Shkoporov, Andrey N.
Publication date	2026-06-01
Original Citation	Bastiaanssen, F T S, Hill, C & Shkoporov, A N 2026, 'Ten quick tips for classifying unknown bacteriophage', PLOS Computational Biology, vol. 22, no. 6, e1014403, pp. 1-9. https://doi.org/10.1371/journal.pcbi.1014403
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.scopus.com/pages/publications/105043672312 - 10.1371/journal.pcbi.1014403
Rights	© 2026, Bastiaanssen et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-10-31 04:01:28
Item downloaded from	https://hdl.handle.net/10468/19055



University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

PLOS Computational Biology

EDUCATION

Ten quick tips for classifying unknown bacteriophage

Fabian T. S. Bastiaanssen*, Colin Hill, Andrey N. Shkoporov

APC Microbiome Ireland and School of Microbiology, University College Cork, Cork, Ireland

* fabian.bastiaanssen@ucc.ie

Check for updates

OPEN ACCESS

Citation: Bastiaanssen FTS, Hill C, Shkoporov AN (2026) Ten quick tips for classifying unknown bacteriophage. PLoS Comput Biol 22(6): e1014403. https://doi.org/10.1371/journal.pcbi.1014403

Editor: Francis Ouellette, Montreal, CANADA

Published: June 23, 2026

Copyright: © 2026 Bastiaanssen et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by Science Foundation Ireland (SFI; https://ror.org/0271asj38) under Grant Numbers SFI/12/RC/2273_P2 and SFI/12/RC/2273 awarded to C.H. This work was also supported by the European Research Council (ERC; https://cordis.europa.eu/project/id/101001684) awarded to A.N.S. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Introduction

In microbiome research, bacteriophages (phages) are gaining increased attention for their roles as important ecological actors, as vehicles of horizontal gene transfer, and as "phagebiotics"—potential tools for precision microbiome manipulation: phage cocktails to suppress specific pathogens/pathobionts, "virome transplants" to restore microbiome diversity and functionality, phage vectors for delivery of CRISPR-Cas for microbiome editing. Once referred to as "viral dark matter" [1,2] of the microbiome, complex populations of bacteriophages are becoming easier to sequence and identify thanks to recent advances in high throughput virome sequencing and phage bioinformatics.

Despite this, characterising complex phage populations and individual genomes continues to be a challenge: too many phages, including isolated ones, lack even a partial taxonomic classification, let alone a complete one. If we discussed bacterial taxonomy the way we do viral taxonomy, we might not even know which bacteria are in the same family as *E. coli*.

This is not due to a lack of effort. Unlike the phylogeny of bacteria, there is not a single clade of bacteriophages that we can trace back to the evolutionary origin of viruses. Viruses have multiple origins and appear to be products of convergent evolution [3], and therefore lack universal marker genes, such as 16S rRNA, that can be used to construct a common phylogenetic tree. To make matters worse, not only are there multiple branches to keep organised, but viruses can also exchange genes with evolutionarily unrelated viral species [4] regardless of their genomic organisation [5] or even the nucleic acid type [6]. A single viral genome can therefore encompass multiple and distinct evolutionary histories. With multiple correct answers possible, taxonomy based on this kind of phylogenetic delineation becomes arbitrary. The International Committee on Taxonomy of Viruses (ICTV), the official authority on viral taxonomy, comprises various expert groups that spend their time discussing classification and delimitation criteria. Anyone can propose adjustments or additions to the current viral taxonomy. The appropriate ICTV subcommittees will then review these proposals and vote on whether to approve them.

Given the challenges involved, it might seem tempting not to bother assigning classifications to our genomes and instead wait for others to figure it all out. The benefit of proactive classification should not be overlooked, however; having a sense of

Competing interests: The authors have declared that no competing interests exist.

taxonomy allows you to link your data to existing literature and databases. While they might be arbitrary, demarcation criteria are, at the end of the day, informed choices that group together similar viruses. If you're interested in the attributes of the phage, you might find similar traits shared by its related taxa. Alternatively, the taxonomy might inform you of traits you were unaware of. If a related phage is extensively studied for its role in phage therapy, you won't overlook similar potential in yours. And most importantly, it makes it easier for others to find (and cite) your work in the future. For those with these noble goals in mind, we present a concise and accessible "reader's digest" in the form of 10 simple tips to guide your path in taxonomically classifying metagenomic and isolated phage genomes (Fig 1). While the principles outlined here apply to phage sequences from any source, we recommend several excellent papers for more specific guidance and further reading [7–9].

Tip 1: Confirm it is actually a virus

Although this may seem like a straightforward tip, it remains critically important. As demonstrated by the recent publication of the Human Gut Archaeal Virome Database (HGAVD) [10], which elicited a response from researchers highlighting numerous bacterial and archaeal false positives and questioning the reliability of CRISPR-based viral detection methods [11]. This ultimately resulted in a final reply wherein the viral assignments of several genomes were revised a third and final time [12].

While this particular case was exaggerated due to the low number of human gut archaeal viruses in reference databases, thereby ironically highlighting the need for the HGAVD, it still shows the importance of confirming whether a genome is actually viral. The use of proven tools such as Genomad [13], VirSorter2 [14], Jaeger [15] is recommended to determine whether a genome is viral and to separate integrated phage regions from host-derived DNA. After all, finding the right viral classification for a non-virus is a fool's errand.

Tip 2: Determine and improve your genome completeness

Due to the small size of viruses and their high level of variability, each uncovered portion of the genome may be of great importance for subsequent analysis steps [16]. It's therefore in our best interest to recover as much of the original genome as possible. CheckV [17] is currently considered the gold standard tool for assessing the completeness and contamination of viral genomes, however, since CheckV relies on machine learning and is constrained by its training data, manual inspection of low and medium-quality genomes is advisable. For instance, if the assembly graph indicates a closed and therefore probably complete, the low quality might be due to CheckV bias rather than low genome coverage. The simplest way to improve coverage is to increase the number of reads in the assembly, either by increasing the read depth or by pooling multiple sequencing runs. Despite being the most prevalent viruses in the global human gut and present in most individuals, the assembly of the first *Crassvirales* [18–20] genomes were only possible through cross-assembly [21]. This is not to say that higher read counts will necessarily lead to higher genome quality; reads should be randomly subsampled to avoid assembly problems caused by excessive coverage (>50-100x) [22].

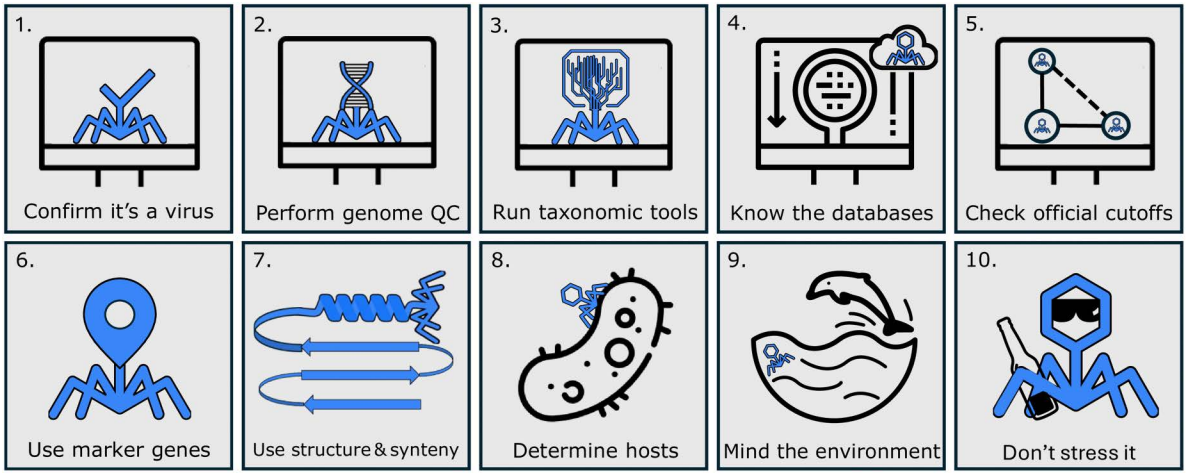


Fig 1. A graphical abstract of the tips.

https://doi.org/10.1371/journal.pcbi.1014403.g001

To maximise the utility of our sequencing data, we can use post-processing tools such as Phables [23] to resolve the assembly graph and improve the size and quality of assembled contigs. Additionally, tools such as vRhyme [24] or PHAMB [25] can cluster viral contigs that likely originate from the same genome into viral metagenome-assembled genomes (vMAGs).

Tip 3: Run a taxonomic tool

Taxonomic classification tools such as taxMyPhage [26], Virsorter2 [14], VIRify [27] and vConTACT3 [28] can help narrow down your classification or even complete it. While these tools may not always determine every taxonomic rank, they will, at the very least, give you a push in the right direction. Each method has inherent strengths and weaknesses in classifying different ranks, which, in combination with the limitations of the reference dataset, often result in partial annotations that are biased toward well-studied viral taxa. Combining multiple tools helps compensate for the method-specific differences and can yield a more comprehensive classification [29]. Importantly, the classifications made by machine-learning-based prediction tools require manual verification and should not be considered definitive. If none of the tools can fully classify your genome, congratulations: you may have discovered a novel taxon or an opportunity to establish new parent taxa for the orphan members.

Tip 4: Curate twice, Blast once

Your genome does not exist in isolation, so don't classify it as if it does. Referencing existing database entries can be very useful. A natural first stop is the ICTV database, which contains gold-standard, officially recognised viral classifications. Finding an exact or close match can allow you to infer taxonomy from existing classifications, while even distant relatives can provide useful context about shared levels of taxonomy. The number of classified phages is much smaller than that of unclassified phages, and it is quite possible that the ICTV contains no close relatives. Even if your genome has no matches for direct taxonomy inference, larger databases such as NCBI GenBank, IMG/VR [30], UHGV [31], INPHARED [32], PHROGs [33], and VIRE [34] might identify close but not formally recognised relatives. These relatives' genomes might prove easier to classify in downstream analyses and, through proxy, inform your phage's classification.

These resources are valuable for identifying sequences, but their taxonomy is unofficial and should be treated cautiously. For example, relying strictly on NCBI's taxonomic labels can inadvertently exclude close relatives listed as "unclassified viruses" or "unclassified bacterial viruses", as these are not grouped by phylogenetic or genomic similarity.

The key is balancing relevance with efficiency: include enough sequences to improve your odds of finding a match, but avoid overloading your analysis with unrelated entries, which slow things down.

Tip 5: Know your cutoffs

The ICTV does not mandate uniform metrics or thresholds. Demarcation criteria are set and enforced for each taxonomic group. Therefore, if you believe your phage belongs to a specific taxon, it is best to review the latest ICTV proposal and determine the current criteria for that taxon. For instance, the ICTV Bacterial and Archaeal Viruses Subcommittee generally recommends approximate cutoffs of ~95% average nucleotide identity (ANI) for species and ~70% for genera [35]. However, this is not universally applicable: for example, members of the order *Crassvirales*, as of MSL #40v2, are classified at the species and genus levels based on the proportion of shared proteins rather than ANI (ICTV Taxonomy proposal 2021.022B.R.Crassvirales). A revision to align *Crassvirales* more closely with the broader viral taxonomy is currently under consideration (ICTV pending proposal 2025.013B.Uc.v4.Crassvirales).

Even when these metrics do not perfectly represent official taxa, having a rough sense of whether your closest matches fall at the species or genus level can substantially improve downstream interpretation. Following the recommended ANI thresholds is particularly helpful during early exploratory analyses, but they must be interpreted with care. Viral genomes vary enormously in size and often share only partially overlapping regions due to extensive horizontal gene transfer. To address this, ANI is sometimes reported alongside coverage, which measures the proportion of the query or reference sequence that aligns with its counterpart, yet this information alone still fails to capture important aspects of relatedness.

For example, take two pairwise comparisons both reporting 95% ANI at 80% coverage. In the first, an 80 kb segment of a 100 kb genome aligns to another 100 kb genome with 95% identity, resulting in roughly 76,000 identical bases and 4,000 mismatches. In the second, a 1 kb genome aligns to 80% of its length (800 bp) in a 100 kb genome at 95% identity, yielding 760 identical bases and 40 mismatches. Although both show 95% ANI and 80% coverage, the actual amount of shared versus divergent sequence differs by nearly two orders of magnitude.

For these reasons, ANI is often more informative when expressed as total ANI (tANI) [9], which accounts for both the coverage of the aligned fraction and alignment length rather than raw identity alone. Tools such as Vclust [36], taxMyPhage [26] and VIRIDIC [35] provide a fast and practical method for calculating ANI, tANI and, related metrics to cluster genomes, providing an initial organisational framework that can be refined later as taxonomy-specific rules are applied.

Tip 6: Find marker genes for phylogenetics

Although no gene is conserved across all viruses, conserved hallmark genes from a common ancestry are found among members of major viral lineages and will be documented in the dedicated ICTV Reports (https://ictv.global/report). These hallmark genes typically encode proteins involved in replication or virion assembly [37–40]. In tailed bacteriophages (*Caudoviricetes*), examples include the major capsid protein (MCP), portal proteins, and the large terminase subunit (terL), all of which play essential roles in virion assembly or DNA packaging and tend to evolve more slowly than accessory genes. terL is functionally constrained and sufficiently conserved to support alignments across large evolutionary distances, making it especially useful for resolving relationships at the family and order levels. Because phage genomes are highly mosaic, relying on a single marker gene can oversimplify their evolutionary history. A multi-gene approach better captures the composite nature of phage genomes while still focusing on their most evolutionarily stable components.

Phylogenetic trees explicitly model evolutionary relationships and help identify nearest neighbours in a broader context. This is particularly valuable when placing a phage into higher taxonomic ranks, such as families or orders, where genome similarity metrics are less informative. At the same time, it is important to recognise the limitations of phylogenetics: trees based on single or a few genes generally struggle to resolve very recent divergences and are therefore poorly suited to fine-scale distinctions at the species or genus level.

In practice, phylogenetic analysis should be viewed as complementary to genome-wide similarity metrics: trees help establish broad evolutionary placement, while ANI, shared protein content, and other quantitative measures refine relationships at lower taxonomic ranks.

Tip 7: Favour structure over sequence

Just as amino acid sequences are more conserved than nucleotide sequences, patterns of genome organisation preserve evolutionary signal even when individual protein sequences have diverged beyond recognition. For example, *Crassvirales* phage often show conserved blocks of genes involved in replication and virion assembly across genera and families, even where sequence similarity fails (Fig 2). Tools such as vConTACT3 [28] and GRAViTy [41] apply this principle to taxonomically cluster phages with greater precision than just identity alone. Similarly, Phynteny [42] can be used to predict the function of unknown genes by analysing surrounding annotations.

Like genome synteny, protein structures evolve more slowly than their underlying sequences, allowing structural similarity to persist long after detectable sequence identity has been lost to sequence erosion [43]. For instance, many phage integrases display high structural similarity to integrases and recombinases from distantly related organisms, despite sharing less than 25% amino acid identity [44]. Structural phylogeny methods, such as FoldTree [45] leverage this conservation to follow the molecular clock more closely than sequence-only trees and to identify evolutionary signals at deeper levels of the tree.

Until recently, the main challenge for structure-based methods was their high resource and time investments. Although tools like AlphaFold2 [46] and ESMFold [47] are much quicker than experimental methods such as cryo-electron microscopy, they still require significant resources, limiting their use in large-scale comparisons. The computationally costly

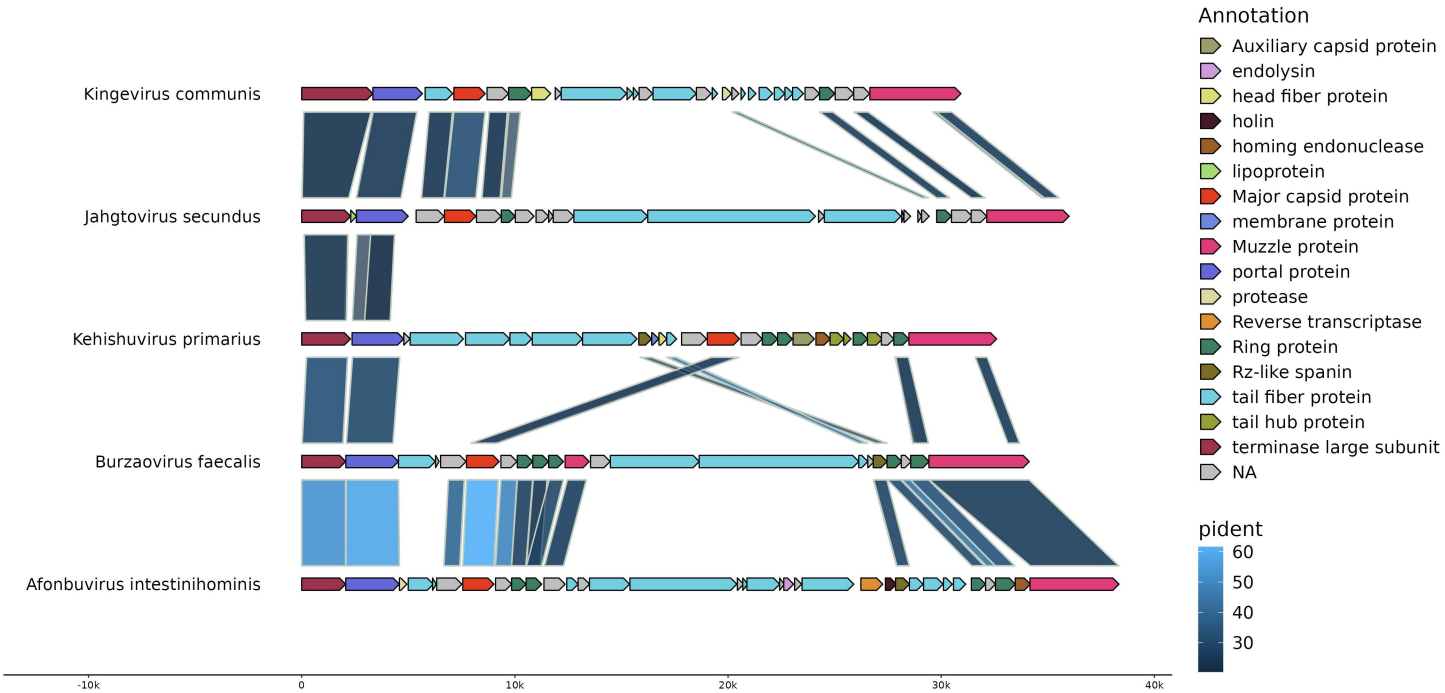


Fig 2. A visualisation of the structural genes within five different members across the families within Crassvirales. Genomes were annotated using Pharokka v1.6 and Prodigal-gv via Pyrodigal, selecting the highest coding density among translation tables 11, 15, and 4. Annotations were further refined using Phold v1 and Phynteny v0.1.12. BLASTp hits (e < 1e-5) between proteins from neighbouring genomes are linked with the brightness of the links scaling with identity.

https://doi.org/10.1371/journal.pcbi.1014403.g002

process of predicting the 3D structure of the protein is mostly wasteful, as the full structure is never used for searches but instead transcribed into a 1D string represented by 20 different characters similar to amino acids, called the 3di alphabet, before any searches and comparisons are possible [48]. ProstT5 [49] tackles this issue by directly predicting 3Di tokens, thereby eliminating the need to first create full 3D models. Benchmark results show that ProstT5 generates 3Di tokens about 50–100 times faster than ESMFold, with only slight decreases in performance [44] and similarity searches using these predicted 3Di tokens can nearly match the sensitivity of structures obtained experimentally and outperform traditional sequence-based methods significantly [49].

Tip 8: Consider potential hosts

As demonstrated by bacteriophages (i.e., phage that infect bacteria), there is clear value in using host range when discussing and categorising viruses. This principle applies across all levels, assigning a phage to a known host genus (e.g., *Escherichia* phage) generally provides more contextual and comparative utility than an alphanumeric isolate ID. In the case of phage isolate genomes, host assignment is usually clear because the host is inherently linked to isolation and propagation. However, for uncultured metagenomes, identifying hosts is notoriously difficult, relies on prediction tools like iPHoP [50], WIsH [51], and PHIST [52], and should not be considered definitive without experimental validation.

However, while the arms race between host and virus shapes the evolutionary history of both parties, host range does not reliably correspond to monophyletic grouping [53,54]. Due to horizontal gene transfer and specialised mutation strategies enabling rapid host switching, a nonlinear relationship exists between phylogeny and host range. Closely related genotypes may differ in phenotypic host range, while more distantly related viruses may share hosts [55].

Shared or similar hosts may therefore point to evolutionary relatedness, but the absence of a shared host should not be taken as strong evidence against it [29]. That said, host range is most useful when sequence similarity alone cannot resolve taxonomy or when it provides relevant biological context for thresholds based on genomic evidence.

Tip 9: Consider the environment

Classifications based on environment, such as “marine viruses” or “gut viromes” are often useful and intuitive for biological and clinical purposes. However, as with host range, they should be applied cautiously to taxonomy, as they can produce polyphyletic groups if evolutionary history and horizontal gene transfer are ignored. Furthermore, since viruses from all realms can be detected in a wide range of environments all across the globe, these categories may actually reflect the ecology of their hosts rather than a constraint on the phage itself [56,57]. As a result, related phages may be found in different environments when their hosts overlap, whereas unrelated viruses can co-occur simply because they occupy the same ecological niche.

That said, at lower taxonomic levels, ecology and taxonomy have been shown to intersect. For example, although distributed worldwide at the order level, within the order *Crassvirales* shows a strong correlation between location and phylogeny, with genetically similar members typically found in close geographic proximity [58]. In practice, this means that a marine phage is more likely to resemble other marine inhabitants within its lineage than those found in soil or the human gut.

Tip 10: Don’t overcomplicate it

At the end of the day, viral taxonomy is a practical tool for conveying shared attributes among related viruses. The choice of ranks and the boundaries between them are human-made constructs for describing the spectrum of viral diversity, which inevitably entails sacrificing some information for utility. Classifications that encompass insights into the biology or ecology of a monophyletic clade are much better than classifications based on thresholds alone.

We should avoid delaying a serviceable classification by attempting to force a “perfect” and exhaustive classification scheme. While additional ranks provide additional information, under the ICTV, the only mandatory ranks are species and

genus, which provide sufficient structure for most purposes. If your analysis reveals interesting features of a species, its ecology, or functional repertoire, a partial but actionable classification for its closest relatives is more useful now than a precise higher-rank placement in the distant future. Further refinement of the classification is not only possible, but inevitable as new species are discovered. Taxonomy is the means to an end, not the end by all means.

Conclusion

Effective viral taxonomy is cumulative, relying on preexisting knowledge and on the rigour with which it is applied to new data. The more rigorously and consistently we classify newly discovered phages, the easier it becomes to place future isolates within a coherent and informative framework. Each well-justified classification strengthens the reference landscape on which all subsequent taxonomic decisions depend.

At the same time, taxonomy is vulnerable to compounding errors, as incorrect or overly confident annotations can spread through databases, bias comparative analyses, and obscure true evolutionary relationships. Because viruses lack fixed, objective thresholds for delimiting taxa, poorly justified classifications can have long-lasting adverse effects on future research. One way to mitigate this risk is to employ multiple methods and compare their classifications. Agreement between methods increases confidence in taxonomic assignments, whereas discrepancies can reveal methodological limitations or biologically interesting edge cases.

When groupings make biological and evolutionary sense, they should be adopted; when they do not, the rules and criteria should be revisited and refined as more data become available. Used thoughtfully, taxonomy is a powerful tool for understanding viruses. Used rigidly, it risks becoming an obstacle rather than an aid.

Lastly, when in doubt, you can always reach out to the ICTV. ICTV members are generally approachable, supportive, and willing to assist researchers by answering questions or directing them toward the most appropriate resources and recommendations.

References

1. Stockdale SR, Ryan FJ, McCann A, Dalmasso M, Ross PR, Hill C. Viral dark matter in the gut virome of elderly humans [Internet]. Preprints; 2018 [cited 2024 Jun 11]. Available from: https://www.preprints.org/manuscript/201807.0128/v1
2. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, et al. Whole-virome analysis sheds light on viral dark matter in Inflammatory Bowel Disease. Cell Host Microbe. 2019;26(6):764–778.e5. https://doi.org/10.1016/j.chom.2019.10.009 PMID: 31757768
3. Nasir A, Romero-Severson E, Claverie JM. Investigating the concept and origin of viruses. Trends Microbiol. 2020;28(12):959–67. https://doi.org/10.1016/j.tim.2020.08.003
4. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017;2:17112. https://doi.org/10.1038/nmicrobiol.2017.112 PMID: 28692019
5. Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, et al. Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. BMC Evol Biol. 2011;11:276. https://doi.org/10.1186/1471-2148-11-276 PMID: 21943216
6. Kazlauskas D, Dayaram A, Kraberger S, Goldstien S, Varsani A, Krupovic M. Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. Virology. 2017;504:114–21. https://doi.org/10.1016/j.virol.2017.02.001 PMID: 28189969
7. Valencia-Toxqui G, Ramsey J. How to introduce a new bacteriophage on the block: a short guide to phage classification. J Virol. 2024;98(10):e01821–23. https://doi.org/10.1128/jvi.01821-23
8. Adriaenssens E, Brister JR. How to name and classify your phage: an informal guide. Viruses. 2017;9(4):70. https://doi.org/10.3390/v9040070 PMID: 28368359
9. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. Viruses. 2021;13(3):506. https://doi.org/10.3390/v13030506 PMID: 33803862
10. Li R, Wang Y, Hu H, Tan Y, Ma Y. Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut. Nat Commun. 2022;13(1):7978. https://doi.org/10.1038/s41467-022-35735-y PMID: 36581612
11. Chibani CM, Shah SA, Schmitz RA, Nayfach S. Inaccurate viral prediction leads to overestimated diversity of the archaeal virome in the human gut. Nat Commun. 2024;15(1):5976. https://doi.org/10.1038/s41467-024-49902-w PMID: 39019907
12. Wang Y, Li R, Ma Y. Reply to: Inaccurate viral prediction leads to overestimated diversity of the archaeal virome in the human gut. Nat Commun. 2024;15(1):5977. https://doi.org/10.1038/s41467-024-49903-9 PMID: 39019854

13. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol. 2024;42(8):1303–12. https://doi.org/10.1038/s41587-023-01953-y PMID: 37735266
14. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9(1):37. https://doi.org/10.1186/s40168-020-00990-y PMID: 33522966
15. Wijesekara Y, Wu LY, Beeloo R, Rozwalak P, Hauptfeld E, Doijad SP, et al. Jaeger: an accurate and fast deep-learning tool to detect bacteriophage sequences. bioRxiv. 2024 [cited 2026 Jan 27]. Available from: https://www.biorxiv.org/content/10.1101/2024.09.24.612722v1
16. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019;37(1):29–37. https://doi.org/10.1038/nbt.4306 PMID: 30556814
17. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39(5):578–85. https://doi.org/10.1038/s41587-020-00774-7 PMID: 33349699
18. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. Nat Commun. 2018;9(1):4781. https://doi.org/10.1038/s41467-018-07225-7 PMID: 30429469
19. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe. 2019;26(4):527–541.e5. https://doi.org/10.1016/j.chom.2019.09.009
20. Guerin E, Shkoporov A, Stockdale SR, Clooney AG, Ryan FJ, Sutton TDS, et al. Biology and taxonomy of crAss-like bacteriophages, the most abundant virus in the human gut. Cell Host Microbe. 2018;24(5):653–664.e6. https://doi.org/10.1016/j.chom.2018.10.002 PMID: 30449316
21. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5:4498. https://doi.org/10.1038/ncomms5498 PMID: 25058116
22. Shen A, Millard A. Phage genome annotation: where to begin and end. Phage. 2021;2(4):183–93. https://doi.org/10.1089/phage.2021.0015 PMID: 36159890
23. Phables: from fragmented assemblies to high-quality bacteriophage genomes | Bioinformatics | Oxford Academic [Internet]. [cited 2026 Jan 27]. Available from: https://academic.oup.com/bioinformatics/article/39/10/btad586/7280146
24. vRhyme enables binning of viral genomes from metagenomes | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2026 Jan 27]. Available from: https://academic.oup.com/nar/article/50/14/e83/6584432
25. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, Deng L, et al. Genome binning of viral entities from bulk metagenomics data. Nat Commun. 2022;13(1):965. https://doi.org/10.1038/s41467-022-28581-5 PMID: 35181661
26. Millard A, Denise R, Lestido M, Nicholas MT, Webster D, Turner D. taxMyPhage: automated taxonomy of dsDNA phage genomes at the genus and species level. PHAGE. 2025;6(1):5–11. https://doi.org/10.1089/phage.2024.0050
27. Rangel-Pineros G, Almeida A, Beracochea M, Sakharova E, Marz M, Muñoz AR. VIRify: An integrated detection, annotation and taxonomic classification pipeline using virus-specific protein profile hidden Markov models. PLOS Comput Biol. 2023;19(8):e1011422. https://doi.org/10.1371/journal.pcbi.1011422
28. Bolduc B, Zablocki O, Turner D, Jang HB, Guo J, Adriaenssens EM, et al. Scalable and systematic hierarchical virus taxonomy with vConTACT3 [Internet]. bioRxiv. 2025 [cited 2025 Nov 21]. Available from: https://www.biorxiv.org/content/10.1101/2025.11.06.686974v1
29. Simmonds P, Adriaenssens EM, Zerbini FM, Abrescia NGA, Aiewsakun P, Alfenas-Zerbini P, et al. Four principles to establish a universal virus taxonomy. PLoS Biol. 2023;21(2):e3001922. https://doi.org/10.1371/journal.pbio.3001922 PMID: 36780432
30. Camargo AP, Nayfach S, Chen IMA, Palaniappan K, Ratner A, Chu K. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Research. 2023;51(D1):D733–43. https://doi.org/10.1093/nar/gkac1037 PMID: 36399502
31. Camargo AP, Baltoumas FA, Ndela EO, Fiamenghi MB, Merrill BD, Carter MM, et al. A genomic atlas of the human gut virome elucidates genetic factors shaping host interactions [Internet]. bioRxiv. 2025 [cited 2026 Feb 24]. Available from: https://www.biorxiv.org/content/10.1101/2025.11.01.686033v1
32. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. Phage (New Rochelle). 2021;2(4):214–23. https://doi.org/10.1089/phage.2021.0007 PMID: 36159887
33. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021;3(3):lqab067. https://doi.org/10.1093/nargab/lqab067 PMID: 34377978
34. VIRE: a metagenome-derived, planetary-scale virome resource with environmental context | Nucleic Acids Research | Oxford Academic [Internet]. [cited 2026 Jan 28]. Available from: https://academic.oup.com/nar/article/54/D1/D902/8356007
35. Moraru C, Varsani A, Kropinski AM. VIRIDIC-a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses. 2020;12(11):1268. https://doi.org/10.3390/v12111268 PMID: 33172115
36. Zielezinski A, Gudyś A, Barylski J, Siminski K, Rozwalak P, Dutilh BE, et al. Ultrafast and accurate sequence alignment and clustering of viral genomes [Internet]. bioRxiv. 2024 [cited 2025 Jan 31]. Available from: https://www.biorxiv.org/content/10.1101/2024.06.27.601020v1

37. Fels JM, Hill AB, Han R, Garcia JM, Bisio H, Abergel C, et al. Giant DNA viruses encode a hallmark translation initiation complex of eukaryotic life. Cell. 2026;189(5):1423–1433.e16. https://doi.org/10.1016/j.cell.2026.01.008 PMID: 41709453
38. Order to the Viral Universe | Journal of Virology [Internet]. [cited 2026 Feb 24]. Available from: https://journals.asm.org/doi/10.1128/jvi.01489-10
39. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N. Global Organization and proposed megataxonomy of the virus world. Microbiol Mol Biol Rev. 2020;84(2) https://doi.org/10.1128/mmbr.00061-19
40. Multiple origins of viral capsid proteins from cellular ancestors | PNAS [Internet]. [cited 2026 Feb 24]. Available from: https://www.pnas.org/doi/full/10.1073/pnas.1621061114
41. Mayne R, Aiewsakun P, Turner D, Adriaenssens EM, Simmonds P. GRAViTy-V2: a grounded viral taxonomy application. NAR Genom Bioinform. 2024;6(4):lqae183. https://doi.org/10.1093/nargab/lqae183 PMID: 39703433
42. Grigson SR, Bouras G, Papudeshi B, Mallawaarachchi V, Roach MR, Decewicz P, et al. Synteny-aware functional annotation of bacteriophage genomes with Phynteny [Internet]. bioRxiv. 2025 [cited 2026 May 11]. Available from: https://www.biorxiv.org/content/10.1101/2025.07.28.667340v2
43. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009;77(3):499–508. https://doi.org/10.1002/prot.22458 PMID: 19507241
44. Bouras G, Grigson SR, Mirdita M, Heinzinger M, Papudeshi B, Mallawaarachchi V, et al. Protein structure-informed bacteriophage genome annotation with Phold. Nucleic Acids Res. 2026;54(1):gkaf1448. https://doi.org/10.1093/nar/gkaf1448 PMID: 41495893
45. Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. Nat Struct Mol Biol. 2025:1–11. https://doi.org/10.1038/s41594-025-01649-8
46. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther. 2023;8(1):115. https://doi.org/10.1038/s41392-023-01381-z PMID: 36918529
47. Evolutionary-scale prediction of atomic-level protein structure with a language model | Science [Internet]. [cited 2026 Feb 2]. Available from: https://www.science.org/doi/10.1126/science.ade2574
48. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42(2):243–6. https://doi.org/10.1038/s41587-023-01773-0 PMID: 37156916
49. Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Steinegger M, Rost B. ProstT5: Bilingual Language Model for Protein Sequence and Structure [Internet]. bioRxiv. 2023 [cited 2024 Jun 14]. Available from: https://www.biorxiv.org/content/10.1101/2023.07.23.550085v1
50. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLoS Biol. 2023;21(4):e3002083. https://doi.org/10.1371/journal.pbio.3002083 PMID: 37083735
51. Galiez C, Siebert M, Enault F, Vincent J, Söding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017;33(19):3113–4. https://doi.org/10.1093/bioinformatics/btx383
52. Zielezinski A, Deorowicz S, Gudyś A. PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. Bioinformatics. 2021;38(5):1447–9. https://doi.org/10.1093/bioinformatics/btab837
53. Shkoporov AN, Turkington CJ, Hill C. Mutualistic interplay between bacteriophages and bacteria in the human gut. Nat Rev Microbiol. 2022;20(12):737–49. https://doi.org/10.1038/s41579-022-00755-4 PMID: 35773472
54. Zinke M, Schröder GF, Lange A. Major tail proteins of bacteriophages of the order Caudovirales. J Biol Chem. 2022;298(1). https://doi.org/10.1016/j.jbc.2021.101472 PMID: 34890646
55. de Jonge PA, Nobrega FL, Brouns SJJ, Dutilh BE. Molecular and evolutionary determinants of bacteriophage host range. Trends Microbiol. 2019;27(1):51–63. https://doi.org/10.1016/j.tim.2018.08.006 PMID: 30181062
56. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A. Marine DNA viral macro- and microdiversity from pole to pole. Cell. 2019;177(5):1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040 PMID: 31031001
57. Chow C-ET, Suttle CA. Biogeography of viruses in the sea. Annu Rev Virol. 2015;2(1):41–66. https://doi.org/10.1146/annurev-virology-031413-085540 PMID: 26958906
58. Edwards RA, Vega AA, Norman HM, Ohaeri M, Levi K, Dinsdale EA, et al. Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. Nat Microbiol. 2019;4(10):1727–36. https://doi.org/10.1038/s41564-019-0494-6 PMID: 31285584