PHAGE: Therapy, Applications, and Research
Volume 2, Number 4, 2021
Mary Ann Liebert, Inc.
DOI: 10.1089/phage.2021.0013

Open camera or QR reader and scan code to access this article and other resources online.

Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation

Dann Turner, PhD,[1,i] Evelien M. Adriaenssens, PhD,[2,ii] Igor Tolstoy, PhD,[3] and Andrew M. Kropinski, PhD[4,5,iii]

Abstract

All sequencing projects of bacteriophages (phages) should seek to report an accurate and comprehensive annotation of their genomes. This article defines 14 questions for those new to phage genomics that should be addressed before submitting a genome sequence to the International Nucleotide Sequence Database Collaboration or writing a publication.

Keywords: bacteriophages, phages, genome annotation, annotation guide, genomics, phage taxonomy

Introduction

Comprehensive and accurate genome annotations and critical assessment of genome completeness are crucial facets for the genome sequencing of all organisms. For bacteriophages, the accuracy of annotation has never been more important. The increasing levels of antibiotic resistance in many bacterial nosocomial pathogens have renewed interest in the exploitation of bacterial viruses as therapeutic[1] and biocontrol agents[2] and in the study of the molecular mechanisms underpinning productive infection.[3] Similarly, our understanding that prophages can influence the fitness, phenotype, and global metabolism of the host lysogen necessitates careful identification and annotation of proviral regions within bacterial genomes.[4,5]

The sequencing of phage genomes allows for the delineation of both close and distant relationships within the wider population of phages. However, for any such assessment to be accurate it needs to rely on the diligent annotation of the genome using both automated methods and manual curation. Annotation is not simply about the identification of open reading frames (ORFs) and the putative function of protein-coding genes but should include, in scope, the identification of other functional elements including transfer RNAs (tRNAs), noncoding RNAs, promoters, and terminators.

Above all, the phage biologist should be aware that errors in assembly can lead to mistakes in annotation that can cause the propagation of inaccuracies in the extant sequence databases.

Metagenomics and viromics methods and analyses rely heavily on high-quality genomic databases and annotations to situate metagenome-derived genomes in sequence space. A small error in a sequence database, whether it concerns the length of a protein, an incomplete genome, or an incorrect functional gene annotation can lead to inaccurate interpretations of rapidly increasing sets of metagenomic data.

[1]Department of Applied Sciences, Faculty of Health and Applied Sciences, University of the West of England, Bristol, United Kingdom.
[2]Quadram Institute Bioscience, Norwich, United Kingdom.
[3]Viral Resources, National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, Maryland, USA.
Departments of [4]Food Science, and [5]Pathobiology, University of Guelph, Guelph, Ontario, Canada.
[i]ORCID ID (https://orcid.org/0000-0002-0249-4513).
[ii]ORCID ID (https://orcid.org/0000-0003-4826-5406).
[iii]ORCID ID (https://orcid.org/0000-0002-6871-6799).

© Dann Turner et al. 2022; Published by Mary Ann Liebert, Inc. This Open Access article is distributed under the terms of the Creative Commons License [CC-BY] (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Correction added on January 21, 2022 after first online publication of December 16, 2021: The article reflects Open Access, with copyright transferring to the author(s), and a Creative Commons License (CC-BY) added (http://creativecommons.org/licenses/by/4.0).

Three of the authors of this article are members of the Bacterial Viruses Subcommittee of the International Committee on Taxonomy of Viruses (ICTV).[6] As such, they are committed to assuring that the sequences of phages submitted to the primary International Nucleotide Sequence Database Collaboration (INSDC) databases (GenBank/EMBL/DDBJ)[7] are of high quality, and that the publications that derive from these submissions are complete and accurate in their annotation and taxonomy. Lamentably, as more people become involved in bacteriophage research, and phage sequencing becomes more high-throughput and automated, we are observing a significant increase in problems.

These include genomes described as circular, chimeric, and incomplete genomes, genomes in which terminal repeats are found in the middle of the sequence, frame shift assembly errors, as well as poorly or incorrectly described gene products.

Herein, we describe a set of questions to help phage neophytes ensure that their genome assemblies and annotations are of sufficient quality to be sustainable in the long term. We cover guidelines for assembly, structural and functional annotation. High-quality, well-annotated genomes are an essential tool for both basic and applied research and they provide the basis for the identification and annotation of related genomes. We describe some of the available tools and appropriate approaches, based on both web graphical user interfaces and the command-line. An overview is presented in Figure 1. For a detailed stepwise command-line approach, est genome assembly, and annotation, we refer the reader to the complementary article ''Phage Genome Annotation: where to begin and end'' by Shen and Millard in this issue.

Question 1: How Was the Genome Sequenced?

DNA sequencing may occur in one's laboratory, in a centralized ''core facility'' or by use of commercial providers. When approaching the latter two, it is important to make them aware of the possibility of terminal redundancies and query the use of Nextera (or other ''tagmentation'') kits for library preparations, which can result in loss of the genome termini (https://phagesdb.org/blog/posts/26).[8,9]

If one has chosen to employ Illumina, Oxford Nanopore Technology (ONT), or PacBio sequencing, we recommend aiming for between 25 and 100× coverage. Hundred- or thousand-fold over-coverage will generally not improve your assembly, is unnecessarily expensive, and may result in assembly errors.[9–12] In comparison to Illumina, there are relatively few reports of phages sequenced solely by ONT or PacBio sequencing. However both ONT and PacBio could be applied for the detection of modified nucleotides or for phages shown to be refractory to conventional sequencing approaches.[13]

Unlike bacterial genomes, where a reference-based assembly can be employed, phage genomes are best assembled *de novo*. A variety of genome assembly software is available for this purpose. In our experience, SPAdes[14] or Shovill (https://github.com/tseemann/shovill) performs well with *de novo* assembly of phage genomes (with Shovill an Illumina-optimized wrapper for SPAdes that includes subsampling procedures). Command-line instructions for the use of SPAdes are provided by Shen and Millard (supplementary protocol). Alternatively, the commercial programs SeqMan from DNASTAR Ultra (Madison, WI) and CLC Workbench (QIAgen) also work well.

All of these programs can also incorporate longer reads from PacBio and ONT sequencing to generate hybrid assemblies; however, for the majority of bacteriophages, long-read technologies and hybrid assemblies will not be required to assemble the complete genome. For a detailed assessment of the application of assemblers on Illumina data at different depths of sequencing, we refer the reader to Rihtman et al.[9] All assemblers that utilize short-read (read Illumina) data inevitably generate low coverage and spurious homopolymeric contigs that should be filtered out.

Assembly metrics should be assessed. How many contigs have been assembled? What is the mean, minimum, and maximum length of the assembled contigs? What is the depth of coverage across the (phage) contig(s) and is it consistent? Differences in coverage between contigs often point toward bacterial contamination, the presence of prophages induced at low levels, and/or the existence of multiple phages in the sample. One should keep in mind that the purification process will influence the presence of host DNA—without a very extended DNase and RNase step, PEG purified preparations generally will include some contaminating host DNA and potentially induced prophages. These assembly statistics can be determined by using programs such as Qualimap[15] and QUAST.[16]

The contig(s) should also be inspected for duplicated regions, which is easily checked for using dotplots (e.g., Gepard[17]) or BLASTN, as contigs may be extended to more than 100% of the phage genome length.

Once the contig has been assessed, a critical step is to map the individual reads back to the genome/contig by using appropriate read alignment software, for example, BWA-MEM,[18] Minimap2,[19] Bowtie2,[20] and samtools.[21] Once done, the mapping should be inspected to identify (1) regions where the paired reads significantly violate the expected distance between paired reads or in the mapped orientation of reads and (2) to identify whether there are any regions of excess or low coverage. Assemblies may require error correction with polishing tools, a process described by Shen and Millard in step 8 of the supplementary protocol.

Question 2: Is the Genome Complete?

The ICTV-acceptable interpretation of the complete genome coverage of bacteriophages is that you have the complete unique sequence of your virus. Phage genomes come in a variety of configurations that have implications for assembly and downstream processing. The most common configurations for isolated tailed phages are circularly permuted (CP), terminally redundant (TR) (defined ends with terminal repeats), or cohesive ends (defined ends with short single-stranded overhangs). Other types are possible, such as ssDNA circular genomes or even dsRNA segmented genomes, which will not be discussed here.

To be considered as complete, in the case of phages that possess TR ends, this means that the sequence includes at minimum one of the redundancies.

PhageTerm[22] accessible as part of the Galaxy Tool Shed[23,24] or via the command-line (see Shen and Millard, step 11 of the supplementary protocol) can be used to provide a prediction of the type of genome termini present, but these predictions should be verified by using run-off Sanger sequencing. For more information of phage termini, we recommend that you consult (Chung et al.[25] and https://phagesdb.org/documents/).

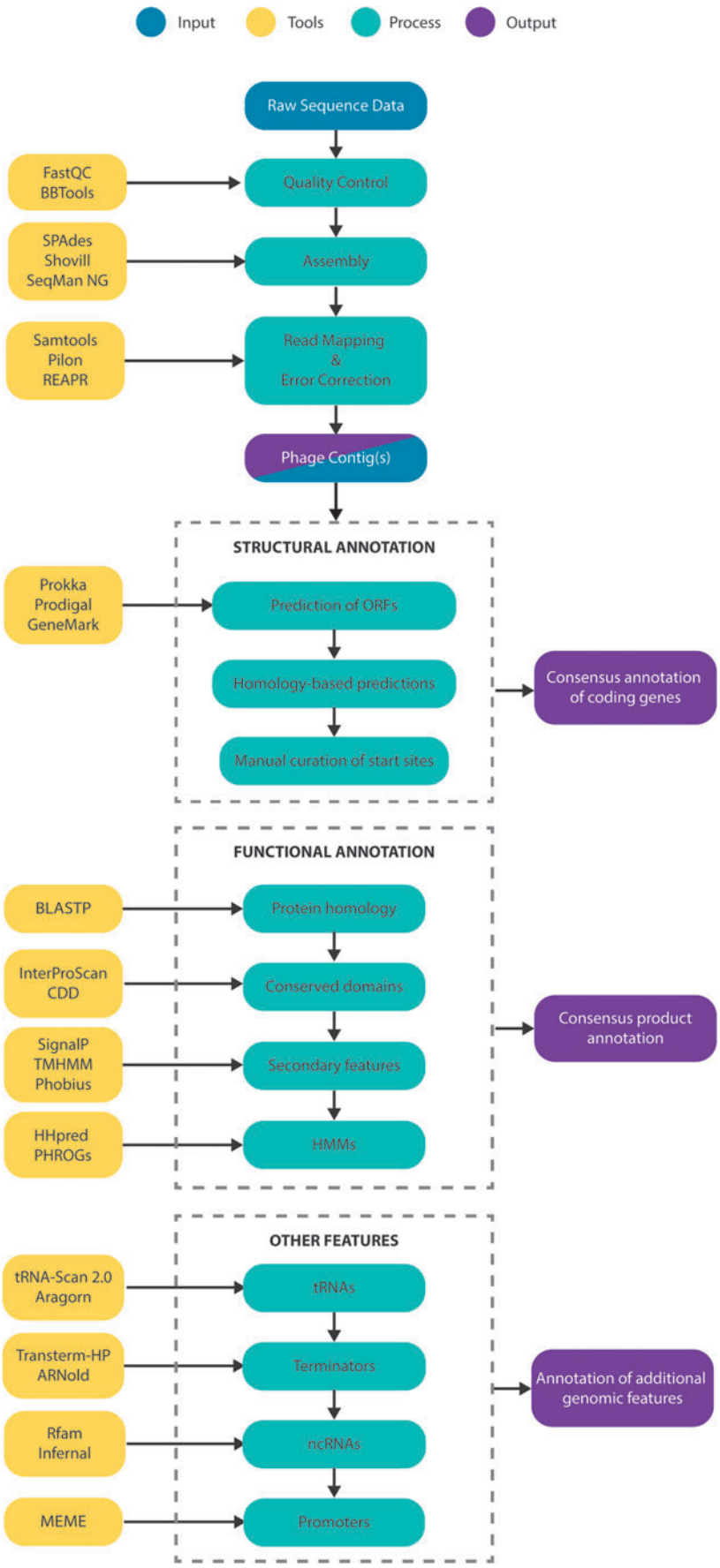

FIG. 1. A recommended workflow for the annotation of structural, functional, and other features in assembled bacteriophage genome sequences. Examples of recommended tools for processes are detailed in yellow, but they do not represent an exhaustive list. CDD, Conserved Domains Database; HMMs, Hidden Markov Models; ncRNAs, noncoding RNAs; ORFs, open reading frames; PHROGs, Prokaryotic virus Remote HOmologous Groups; tRNAs, transfer RNAs.

Question 3: Does the Sequence Contain Ambiguous Bases or Potential Frameshifts?

Ambiguous bases, often denoted by the International Union of Pure and Applied Chemistry (IUPAC) code of N (any base), should be avoided. These discrepancies should be resolved by targeted Sanger sequencing and can be identified by any online resources that count bases.

Three varieties of frameshifts exist—sequencing/assembly errors, programmed frameshifts, and introns. Should a homologous phage exist then at the DNA level BLASTX can often be used to identify frameshifts. Programmed frameshift is often discovered as a result of in-depth proteomic analyses. In the case of coliphage lambda, tail assembly presents an example of a programmed translational frameshift, resulting in the gpGT protein,[26] a feature also uncovered in tails of other bacteriophages.[27,28] Another common example is the presence of two major capsid proteins (10A/10B) in *Escherichia* phage T7.[29]

Question 4: Is the Genome Assembly Co-linear with That of Closely Related Genomes?

Circularization of phage genomes is a common artefact of assembly. No dsDNA tailed phage genome is truly circular when packaged in the capsid, but it can be CP or have terminal repeats that result in an *apparently* circular assembly. With CP genomes it is often easy to convince oneself that the genome is circular[30] but this is not the case. In addition, during the assembly of TR genomes, the redundancies may end up being located internally within the contig (Fig. 2).

If your phage genome is CP, then the choice of ''beginning'' and ''end'' is an arbitrary one. If your phage is related to one already present in the National Center for Biotechnology Information (NCBI), then your chromosome should be made co-linear with that of the reference genome or the ''type virus.'' This can be readily ascertained from the BLASTN output or using a visualization tool such as progressiveMauve[31] or clinker (https://github.com/gamcil/clinker).

Without a homologous genome in the database, and without defined ends, we suggest that you choose the ''beginning'' in an intergenic region that separates operons. Inspiration can be found by using more distant relatives, for example after the *rIIB* gene for T4-like phages.[32] For siphoviruses, past convention has been to open the assembly at the start codon of the small terminase subunit.

By ensuring that the genome begins at experimentally defined termini, or opening the genome at a specific gene, one avoids the common problem of an ORF being split across the ends of the contig and enables easier interpretation of visual pairwise comparison of genomes using visualization tools.

Methods for the identification of genome termini using PhageTerm, and the reordering of phage genomes are detailed in Shen and Millard (supplementary protocol, steps 11 and 12).

Question 5: How Was the Genome Annotated?

Annotation can be sub-divided into structural annotation, the identification of coding sequences (CDSs), and functional annotation, the identification of gene products. Here, we will distinguish between annotation with visual oversight (manual annotation) and without oversight (auto-annotation). Although the SEA-PHAGES (Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science; https://seaphages.org/) program employs the Windows-based program DNA Master[33] to annotate bacteriophage genomes, many use auto-annotation as the first step toward the complete analysis of our phage genomes.

Unfortunately, most of the programs that we use were designed for the annotation of bacterial genomes that tend to possess larger ORFs than those of their viruses. This

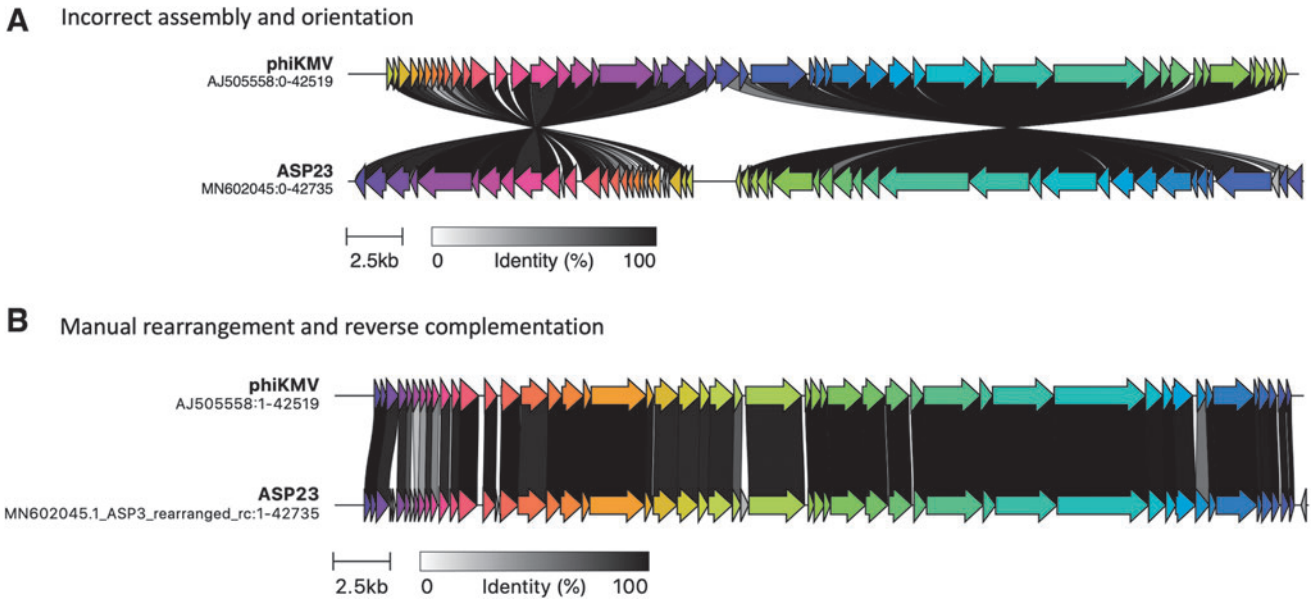

FIG. 2. Genome CDS comparison between Pseudomonas phages phiKMV (AJ505558) and vB_PaeP_ASP23, short ASP23 (MN602045) using clinker. Homologous CDSs are in the same color and linked through gray bars with the percentage amino acid identity, as indicated in the legend. **(A)** Direct comparison and visualization of GenBank records (accessed October 25, 2021). **(B)** Manual rearrangement of the ASP23 genome in TextEdit on Mac and reverse complementation with the Sequence Manipulation Suite,[T14] followed by reannotation with Prokka.[89] CDS, coding sequence.

Table 1. Software for Phage Genome Annotation

Program	*Usage*	*Source*	*Reference*
Prokka	MAC, Linux/Galaxy	https://github.com/tseemann/prokka	89
RAST	WEB	https://rast.nmpdr.org/	90,91
PATRIC	WEB (RAST)	https://www.patricbrc.org/	92
DFAST	WEB	https://dfast.ddbj.nig.ac.jp/	93,94
Phage-specific Galaxy instance	WEB	https://cpt.tamu.edu/	95
UGENE	WIN, MAC, Linux	http://ugene.net/	35
Phage Commander	MAC, Linux	https://github.com/mlazeroff/PhageCommander	49
multiPhATE	MAC, Linux/Galaxy	https://github.com/carolzhou/multiPhATE2	96
DNA Master	WIN	https://phagesdb.org/DNAMaster/	33
DNASTAR (COM)	WIN, MAC	https://www.dnastar.com/	97
Geneious (COM)	WIN, MAC	https://www.geneious.com/	98
VIGA	MAC, Linux	https://github.com/EGTortuero/viga	

COM, commercial software; MAC, Mac computer; WEB, Internet resource; WIN, Windows.

means that many, particularly smaller, CDSs may have been missed. To rectify this problem, one needs freeware such as Artemis[34] (https://www.sanger.ac.uk/tool/artemis/), DNA Master, UGENE,[35] or commercial programs such as DNASTAR or Geneious (Table 1) to examine the DNA sequence for ORFs and potential CDSs, considering all ORFs >75 nt.[36]

Phage genomes possess a high coding capacity; therefore, one might expect to see small gaps between CDSs or small overlaps. Large gaps should immediately alert you to potentially missing genes and should be manually verified, though some phages such as *Escherichia* phage rV5[37] do possess wastelands with no apparent genes.

There are three unusual cases that the researcher should keep in mind while annotating their genomes. These include CDSs within CDSs, frameshifts (e.g., GpG, GpGT tail assembly chaperones, as mentioned earlier), and the presence of introns and inteins. Although embedded genes are rare among members of the class *Caudoviricetes*, there are some notable instances where they can be found.[38–40] One of the best examples is in the lysis cassette of Escherichia virus λ where genes *S*, *R*, *Rz*, and *Rz1* encode the holin, endolysin, i-spannin, and o-spannin, respectively (Fig. 3). We recommend that you employ extreme caution in recording embedded genes in your genome annotation.

The presence of inteins[41] or introns is usually suspected if the size of the gene is significantly different from that of the expected protein, or if the gene is split across multiple ORFs (Fig. 4). Introns are relatively common in phages with large genomes, particularly those that infect *Campylobacter* and *Staphylococcus*, but size should not be used as an excuse not to examine for their presence since they have been found in many differently sized viruses.[42–45] The intron possesses ribozyme activity, which splices itself out of the messenger RNA (mRNA), resulting in the mature mRNA, which is translated into a protein. This region may encode a homing endonuclease, which is the case with *Lactobacillus* phage LL-H (Fig. 4A).

FIG. 3. An example of an embedded gene—Rz1 within R of phage lambda, adapted from Rajaure et al.[115]

Including all possible internal and overlapping ORFs as CDSs can lead to over-annotation of a genome (Fig. 5). In the case of phage Felix01, the originally annotated sequence of all the genes corresponded to 113.7 kb, that is, a coding capacity of 1.32. When the annotation was curated as part of the NCBI Reference Sequence validation process,[46] the coding capacity was reduced to 0.90 (77.6 kb) after the removal of spurious ORFs—a value that is near the norm for phages belonging to the class *Caudoviricetes*.

Question 6: Were CDS Start Sites Curated?

Most tailed phages use the general bacterial translation table (table 11), with ATG and GTG being the most common start codon, but also a set of alternative initiation codons (TTG, CTG, ATT, and ATC).[47,48] These alternative start codons are not always recognized by automated gene calling software, and differences in gene calling between programs can now be assessed with the new tool PhageCommander.[49] Best practices, therefore, include the manual curation of the start codons to optimize coding density and ensure that ribosome binding sites (RBS, or Shine–Dalgarno sequences) precede each start codon.

A meta-analysis of all known viruses showed the presence of an RBS in the form of AGGAGG (or 4mer/5mer variations of this sequence) in more than 50% of the CDSs of all bacteriophages.[50] One way of assessing the presence or absence of ribosome binding sites is to extract sequences that encompass 100 bp upstream of the ORF and the predicted start codon for analysis using MEME[51] (Fig. 6).

Question 7: How Were the Functions of the Gene Products Identified?

There are now a wealth of tools to enrich the functional annotation of gene products beyond that obtained by BLASTP alone, which in many cases appears to be the only tool used. Moreover, an over-reliance on automated BLASTP functional assignments can result in the miss-annotation of proteins, a problem that has the potential to propagate errors through the INSDC databases. An appropriate approach is to look for some manner of consensus across a number of different annotation methods. In addition to BLASTP and Position Specific Iterated BLAST (PSI-BLAST), we recommend the use of InterPro,[52] Batch Web CD-Search Tool,[53] and

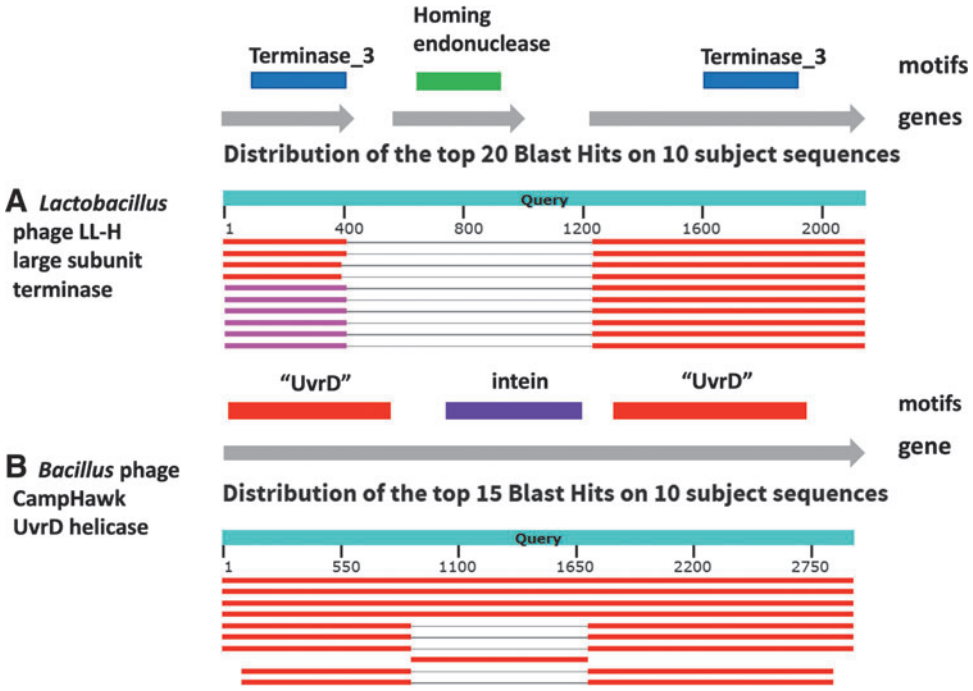

FIG. 4. **(A)** BLASTX analysis of the Lactobacillus phage LL-H large subunit terminase gene (NC_009554.1, 1647..3785) in which it appears that this region contains three ORFs, two of which on translation show homology to TerL proteins and one similarity to a homing endonuclease. **(B)** BLASTX analysis of Bacillus phage CampHawk (NC_022761.1, 118288..121224), which encodes a UvrD-like helicase with an intein.[116] A nearly identical sequence is found in Bacillus phage vB_BsuM-Goe2, but not in Bacillus phage SP8. The CampHawk helicase contains 978 amino acids, whereas its homolog in SP8 possesses only 703 residues.

HHpred (or the command-line instance HHsuite).[54] Another useful sequence feature to scan for are transmembrane domains, which can be readily identified by using Phobius[55] and TMHMM.[56]

Please beware of calling a protein a DNA polymerase based on no or poor quality evidence since this leads to miss-annotation creep and database poisoning, which is far worse than designating a DNA polymerase as a ''hypothetical protein.'' Again please exercise caution by not over-relying on BLASTP hits.

In addition to the recommendations outlined here, Shen and Millard provide routes to customize the automated annotation of gene products with Prokka that implement the Prokaryotic virus Remote HOmologous Groups (PHROGS) and custom-tailed phage databases for improved inference of function (supplementary protocol, steps 14 and 15).

A note on standardized terminology

One of the most common proteins encoded by T4-like phages is RIIA, yet an examination of homologs in NCBI reveals names that vary from ''rIIA protector from prophage-induced early lysis,'' ''membrane-associated affects host membrane ATPase,'' and ''rIIA protein,'' to ''hypothetical protein,'' ''unnamed protein product,'' and ''protein of unknown function.'' The last three annotations are examples of poor annotation since the functions of homologs of this protein are clearly known.

Please note that ''phage hypothetical protein'' is redundant since all of your viruses' proteins are ''phage proteins.'' If no function can be predicted, the term ''hypothetical protein'' should be used as the product qualifier value.

Product names such as ''UboA,'' ''Mcp,'' ''hypothetical protein SA5_0153/152,'' ''ORF184,'' ''gp200,'' ''RNAP1,'' ''32 kDa protein,'' and ''hypothetical protein HY02_082'' should not occur in your annotated genome because they do not mean anything to the casual, or indeed even the informed, reader. The use of ''gp'' (gene product) is common but should be discouraged since gp200 describes radically different proteins in *Listeria*, *Enterococcus*, *Mycobacterium*, *Rhodococcus*, *Sphingomonas*, *Pseudomonas*, *Bacillus*, and *Synechococcus* phage genomes. If you want to relate it to an existing protein, you can add the information as a note to the annotation, for example, /note = ''similar to gp43 of Escherichia phage T4.''

Unless you are a bioinformatician or biostatistician, or an expert in a specific protein family, be very conservative in recording a protein function from ''BLAST hits.'' Could you convince your grandmother? If not, list it as a ''hypothetical protein.'' If you have some homology or motif data that suggest that it may be a DNA polymerase, describe it as a ''hypothetical protein'' and then add an ''evidence qualifier'' (see, https://www.ncbi.nlm.nih.gov/genbank/evidence/) such as /inference = ''protein motif DNA_pol_B_2 (PF03175).'' Do not name the gene product ''putative DNA polymerase.''

We would recommend that you consult the UniProt Knowledgebase (UniProtKB[57]), which is a manually curated and reviewed information database on proteins (https://www.uniprot.org/) and ViralZone[58] (https://viralzone.expasy.org/) when assigning names to gene products.

Question 8: How Did You Screen for Integrases/Recombinases, Toxins, and Antibiotic-Resistance Genes?

Since many phages are isolated as potential therapeutic agents, the presence of indicators of a temperate lifestyle and the carriage of toxin genes would preclude their use.

Predictions of temperate lifestyles from genomic data were traditionally done by manual scanning of the predicted proteins for lysogeny-related genes (integrases/recombinases/transposases). Automated lifestyle predictions can be done with PhageAI,[59] PHACTS,[60] or BACPHLIP,[61] which are based on the presence of conserved domains, or in the case

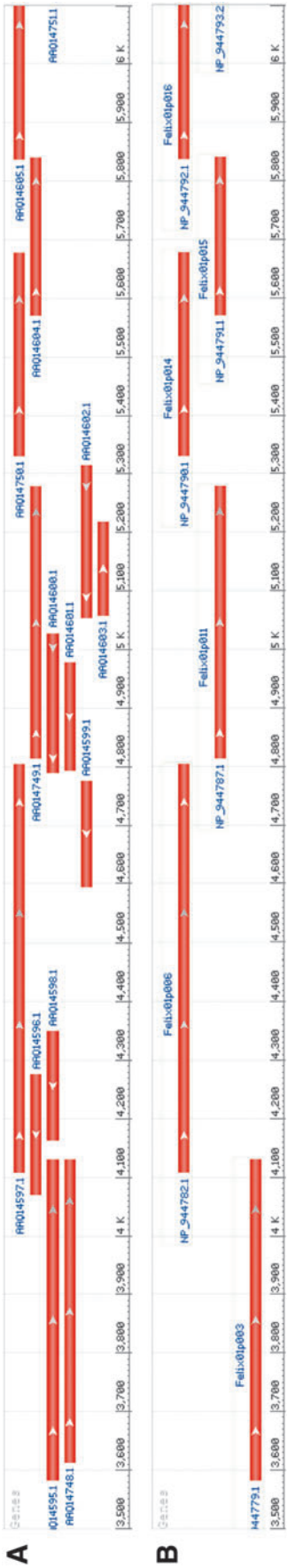

FIG. 5. An example of an over-annotated genome. **(A)** Bacteriophage Felix 01 GenBank record (AF320576), graphical view of the region 3500..6100. **(B)** RefSeq curated record for Enterobacteria phage Felix 01 (NC_005282.1) graphical view of the region 3500..6100. Annotated coding sequences are depicted as red rectangles, with arrows denoting presence on the forward or reverse strand.

of phage.ai through machine learning and natural language processing. As with all automated predictions, use caution when the evidence scores are low.

Although there is plenty of evidence that bacterial antibiotic-resistance genes (ARGs) genes can be disseminated by phage-mediated transduction,[62] and indeed viromes have been demonstrated to have associated ARGs,[63–66] the identification of these elements in phage genomes should be treated with extreme caution and a rare occurrence.[67] The Comprehensive Antibiotic Resistance Database[68] (https://card.mcmaster.ca) is a vital resource, but the results should be treated with great skepticism unless predicted functions have been experimentally verified. Another resource for the identification of ARGs is AMRFinderPlus, developed at the NCBI.[69]

The presence of toxin-encoding genes in the phage genome immediately precludes the use of that phage for therapeutic purposes. A number of Internet resources will assist you in determining whether your phage carries a toxin gene (Table 2). Once again use caution when interpreting the results.

Further considerations for requirements for the annotation of therapeutic phages are discussed by Shen and Millard.

Question 9: Were Putative Promoters, Terminators, and Other Elements Identified?

Though not strictly necessary in genome submissions or publications, some authors choose to screen their genomes for host- or phage-RNA polymerase-dependent promoters, and ρ-independent terminators. To our knowledge, no one has yet analyzed ρ-dependent terminators.[70] Since, in the absence of RNA-seq data, these elements are all theoretical, we recommend that authors err on the side of caution when presenting data. Toward this end, promoters and terminators should only be added to the annotation if they are at the 5′ or 3′-end of genes, respectively, or in the intergenic regions.

Promoters employing RpoD (Sigma70)-dependent host RNA polymerase can be identified by using a variety of online sites (Table 3).

The host housekeeping (RpoD-dependent) promoters in Gammaproteobacteria possess the consensus sequence TTGACA[N15-19]TATAAT.[71] We recommend that you consult the literature for the latest consensus sequence for the bacterium of interest.[72] One of the authors (A.M.K.) only records sequences that differ by two nucleotides or less from the consensus. It is appropriate, but not obligatory, to include the transcriptional start site (+1). If your putative promoter includes an AT-rich upstream promoter DNA element,[73] please include it. If not, trim to the consensus.

Factor ρ-independent terminators can be identified by using a variety of online tools (Table 4) and should be trimmed to remove sequences 5′ of the uploop and 3′ of the final thymidylate residue. In this case, we recommend that you only record terminators that display a Gibbs free energy (ΔG) equal to or lower than −10 kcal/mol.[74]

Lastly, tRNAs are a common structural element, particularly in large phage genomes.[75] The two programs that we recommend are tRNAscan-SE at http://lowelab.ucsc.edu/tRNAscan-SE/[76] and ARAGORN at www.ansikte.se/ARAGORN/.[77] Please note that bacterial and bacteriophage tRNAs possess a CAA triplet at their 3′-termini, which are sometimes missing from auto-annotated genomes.

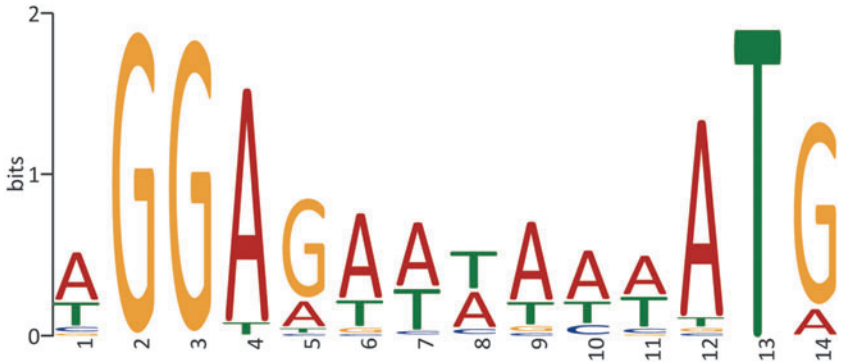

FIG. 6. Sequence logo of a statistically over-represented motif (Shine–Dalgarno sequence or ribosome-binding site) identified by using MEME from 103 bp sequences encompassing the predicted start codon and upstream region.

Question 10: Does the Sequence Represent a Phage Isolate, Prophage, or a Metagenome-Derived "Envirophage"?

Unless specified within the annotation submitted to the INSDC, it is often impossible for the reader to discern whether a deposited sequence represents a lytic or temperate phage isolated by using a specific host, an induced prophage only known from its bacterial genome coordinates, or a sequence derived from metagenomic analyses (envirophage). Before submission, we recommend that phage biologists look in detail at the available source feature keys (https://www.insdc.org/documents/feature-table#7.3).

The qualifier ''/proviral'' can be used within the source feature key to denote that the sequence has been obtained from an induced prophage using the bacterial genome sequencing data. In addition, the qualifier ''/host'' can be used to inform the reader which host strain a prophage was induced from. The qualifier ''/lab_host'' can be used to denote the host stain used for isolation and/or propagation. Additional information, for example ''/isolation_source'' or ''/country,'' could also be made included to expand the available information.

Similarly, for ''envirophages,'' the qualifier/environmental_sample indicates that the genome sequence was derived, not from an isolate, but from a bulk environmental nucleic acid sample, without culturing. This qualifier should always be used in conjunction with the /isolation_source qualifier to indicate the type of sample the sequence was derived from. Phage genomes with these qualifiers will get an ''ENV'' tag in databases.

Question 11: Have You Chosen a Realistic and Useful Phage Name and Locus Tag?

Here, we must reintroduce the importance of choosing a ''good'' name for your phage. We strongly recommend that authors consult Adriaenssens and Brister[78] together with NCBI (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/), Phage Name Check (www.phage.org/phage_name_check.html), and CPT Phage Name Search (https://cpt.tamu.edu/phage-registry/) to assure that the phage name is unique. Although simple alphanumerically names such as P22 and T4 used to be the norm, today most people are following the lead of the SEA-PHAGES program[79,80] and opt for a common name such as Abrogate, Adjutor, and Adonis.

These names greatly assist in the creation of new genera, and binomial species names,[81] which often use the unique part of exemplar phage names for inspiration. In addition, ''Locus_tags are identifiers that are systematically applied to every gene in a genome. These tags have become surrogate gene names by the biological community.'' (https://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf) We recommend that you employ the unique part of the phage name as part of the locus tag. In addition, if your group plans on submitting many different sequences it may be worth considering adding a common set of letters to the locus tag for recognition purposes.

Question 12: Does the Submitted Sequence Contain Irrelevant Material?

Before submission to one of the three major INSDC databases, you want to clean up your submission to remove irrelevant material. Please check that the genome is indicated as linear, not circular. The ''Definition'' line should read ''host genus + phage + name,'' for example, Proteus phage Mydo. Do not include host species name, isolate, chromosome, DNA, or genome. Many of the automated annotation programs leave evidence of their use, such as RAST ''db_xref = "SEED:fig|6666666.271554.peg.1''

All local/temporary identifiers that cannot be found by external users on the web should be deleted. This can be easily accomplished in using a text editor such as Notepad (Windows), TextEdit (Mac), or the more powerful freeware tool Notepad++ (Windows).

TABLE 2. INTERNET RESOURCES FOR TOXIN SCREENING

Program name	*URL*	*Comment*	*Reference*
VirulenceFinder 2.0	https://cge.cbs.dtu.dk/services/VirulenceFinder/	Not all pathogens	99
VirulentPred	http://203.92.44.117/virulent/submit.html		100
DBETH	www.hpppi.iicb.res.in/btox/cgi-bin2/svm.cgi?name=svm	Single sequence	101
T3DB	www.t3db.ca/biodb/search/target_bonds/sequence	Single sequence	102
VFDB	www.mgc.ac.cn/VFs/search_VFs.htm	Single sequence	103

Table 3. Internet Resources for Screening Your Phage Genome for Promoters

Program name	*URL*	*Comment*	*Reference*
Bacterial promoters			
PromoterHunter	www.phisite.org/main/index.php?nav=tools&nav_sel=hunter	Part of phiSITE	104
PhagePromoter	https://galaxy.bio.di.uminho.pt/?tool_id=get_proms &version=0.1.0		105
Genome2D (Prokaryote Promoter Prediction)	http://genome2d.molgenrug.nl/g2d_pepper_promoters.php	Part of PePPER	106
SAPPHIRE	https://sapphire.biw.kuleuven.be/		107
Phage promoters			
PHIRE	https://www.biw.kuleuven.be/logt/PHIRE.htm	Very slow; WIN	108
MEME	https://meme-suite.org/meme/	Part of the MEME Suite	51
STREME	https://meme-suite.org/meme/tools/streme	Part of the MEME Suite	109

Question 13: Is the Taxonomy That I Have Assigned to This Phage Correct?

Although a detailed discussion of phage taxonomy is outside of the scope of this article, some general thresholds can be reiterated.[82] If your phage exhibits $\geq$95% sequence similarity (by VIRIDIC[83] or for BLASTN,[84] coverage multiplied by identity), then it represents a new strain of an existing phage species. If it exhibits <95% but $\geq$70% sequence similarity it represents the first isolate of a new species in an undefined or existing genus. The delineation of subfamily or family-level relationships requires more careful inspection, including pangenome analysis and the inference of single gene or concatenated/partitioned signature gene phylogenies. In the case of confusion or questions, please consult the appropriate member of the Bacterial Viruses Subcommittee of ICTV (https://talk.ictvonline.org/information/members-606089945/w/members/441/bacterial-viruses-subcommittee).

Question 14: Should I Update My Database Submission When New Data Renders It Dated?

Data do not rest, nor does the phage research community. Alongside homology searches of CDSs, it is worthwhile reading the published literature associated with closely related phages. Often, this will reveal experimental evidence regarding the identification of phage structural proteins, such as from mass spectrometry data. Similarly, any studies of phage proteins encompassing cryo-electron microscopy, crystallography, or nuclear magnetic resonance can enrich your annotation given an appropriate level of sequence homology and predicted protein secondary structure. Transcriptomics studies of phage infection may lead to updated information about start codons and location of regulatory elements.

It is important to remember that you ''own'' your genome sequence and annotation; the INSDC prefers that any updates arise from the original submitter and as such the authors should take responsibility for updating their own records. Please note, that any phage sequence submitted without annotation is automatically marked as ''unverified'' according to GenBank policy.

Final Statement

Sequencing, assembling, and annotating a newly isolated phage is a rewarding process and contributes data to our knowledge of the global phage population. However, the value and utility of these data is dependent on a careful, measured, and diligent approach to these processes. It is our hope that the answers to these questions will provide direction and prove useful to the wider community of phage biologists. Similarly, the tools and programs detailed here do not represent an exhaustive list of those available, and new tools are developed all the time. We highly recommend consulting Shen and Millard, who provide additional considerations on these 14 questions as well as a step-by-step guide to phage genome assembly and annotation that employs a number of tools mentioned here.

Happy annotating!

Further Practical Information Can Be Found In…

Bacteriophages: Methods and Protocols Volume 3[85]:

- Chapter 9: Sequencing, Assembling, and Finishing Complete Bacteriophage Genomes[86]

Table 4. Internet Resources for Screening Your Phage Genome for ρ-Independent Terminators

Program name	*URL*	*Comment*	*Reference*
Genome2D (Transcription Terminator Prediction)	http://genome2d.molgenrug.nl/g2d_pepper_transterm.php	Part of PePPER	106
ARNold	http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/	Nice output	110
FindTerm	www.softberry.com/berry.phtml?topic=findterm&group= programs&subgroup=gfindb	Part of SoftBerry suite	111
iTerm-PseKNC	http://lin-group.cn/server/iTerm-PseKNC/		112,113

- Chapter 11: Analyzing Genome Termini of Bacteriophage Through High-Throughput Sequencing[87]
- Chapter 16: Annotation of Bacteriophage Genome Sequences Using DNA Master: An Overview[33]
- Chapter 18: Visualization of Phage Genomic Data: Comparative Genomics and Publication-Quality Diagrams[88]

We encourage users to read publications associated with the tools and programs mentioned here and to use the variety of discussion-board platforms available (e.g., SeqAnswers, Biostar, ResearchGate) to search for advice and for trouble-shooting.

Authors' Contributions

D.T., E.M.A., and A.M.K.: conceptualization, analyses, drafting of the article, and final editing. I.T.: analyses, final editing. All authors approve the final version of the article.

Author Disclosure Statement

The authors declare no conflict of interest.

Funding Information

E.M.A. gratefully acknowledges 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. The work of I.T. was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

References

1. Gordillo Altamirano FL, Barr JJ. Phage therapy in the postantibiotic era. Clin Microbiol Rev. 2019;32(2):e00066-18.
2. Holtappels D, Fortuna K, Lavigne R, et al. The future of phage biocontrol in integrated plant protection for sustainable crop production. Curr Opin Biotechnol. 2021;68:60–71.
3. De Smet J, Hendrix H, Blasdel BG, et al. Pseudomonas predators: Understanding and exploiting phage–host interactions. Nat Rev Microbiol. 2017;15(9):517–530.
4. Moreno Switt AI, Sulakvelidze A, Wiedmann M, et al. Salmonella phages and prophages: Genomics, taxonomy, and applied aspects. In: Schatten H, Eisenstark A; eds. *Salmonella: Methods and Protocols, Methods in Molecular Biology: Volume 1225*. New York: Springer Science + Business Media; 2015: 237–287.
5. Christie GE, Kuzio HE, McShan J, et al. Prophage-induced changes in cellular cytochemistry and virulence. In: Hyman P, Abedon ST; eds. *Bacteriophages in Health and Disease*. CABI Press; 2012.
6. Krupovic M, Turner D, Morozova V, et al. Bacterial viruses subcommittee and archaeal viruses subcommittee of the ICTV: Update of taxonomy changes in 2021. Arch Virol. 2021;166(11):3239–3244.
7. Karsch-Mizrachi I, Nakamura Y, Cochrane G. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2012;40(D1):D33–D37.
8. Roux S, Solonenko NE, Dang VT, et al. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. Peer J. 2016;4:e2777.
9. Rihtman B, Meaden S, Clokie MRJ, et al. Assessing illumina technology for the high-throughput sequencing of bacteriophage genomes. Peer J. 2016;4:e2055.
10. Pightling AW, Petronella N, Pagotto F. Choice of reference sequence and assembler for alignment of listeria monocytogenes short-read sequence data greatly influences rates of error in SNP analyses. PLoS One. 2014;9(8):e104579.
11. Desai A, Marwah VS, Yadav A, et al. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. PLoS One. 2013;8(4):e60204.
12. Wang Y, Yu Y, Pan B, et al. Optimizing hybrid assembly of next-generation sequence data from Enterococcus faecium: A microbe with highly divergent genome. BMC Syst Biol. 2012;6(S3):S21.
13. Lu S, Le S, Tan Y, et al. Unlocking the mystery of the hard-to-sequence phage genome: PaP1 methylome and bacterial immunity. BMC Genomics. 2014;15(1):803.
14. Nurk S, Bankevich A, Antipov D, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In: Deng M, Jiang R, Sun F, Zhang X; eds. *Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer Science: Volume 7821*. Berlin, Heidelberg: Springer; 2013: 158–170.
15. García-Alcalde F, Okonechnikov K, Carbonell J, et al. Qualimap: Evaluating next-generation sequencing alignment data. Bioinformatics. 2012;28(20):2678–2679.
16. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–1075.
17. Krumsiek J, Arnold R, Rattei T. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007;23(8):1026–1028.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013;arXiv:1303.3997 [q-bio.GN].
19. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–3100.
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–359.
21. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2):giab008.
22. Garneau JR, Depardieu F, Fortier L-C, et al. PhageTerm: A tool for fast and accurate determination of phage termini and packaging mechanism using next-generation sequencing data. Sci Rep. 2017;7(1):8292.
23. Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46(W1):W537–W544.
24. Mareuil F, Doppelt-Azeroual O, Ménager H. A public Galaxy platform at Pasteur used as an execution engine for web services. F1000 Res. 2017;6:1030 (poster).
25. Chung CH, Walter MH, Yang L, et al. Predicting genome terminus sequences of Bacillus cereus-group bacteriophage using next generation sequencing data. BMC Genomics. 2017;18(1):350.
26. Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell. 2004;16(1):11–21.
27. Dorscht J, Klumpp J, Bielmann R, et al. Comparative genome analysis of listeria bacteriophages reveals extensive mosaicism, programmed translational frameshifting,

and a novel prophage insertion site. J Bacteriol. 2009; 191(23):7206–7215.
28. Olia AS, Cingolani G. A shifty stop for a hairy tail. Mol Microbiol. 2008;70(3):549–553.
29. Condron BG, Atkins JF, Gesteland RF. Frameshifting in gene 10 of bacteriophage T7. J Bacteriol. 1991;173(21): 6998–7003.
30. Denyes JM, Krell PJ, Manderville RA, et al. The genome and proteome of Serratia bacteriophage η which forms unstable lysogens. Virol J. 2014;11(1):6.
31. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
32. Miller ES, Kutter E, Mosig G, et al. Bacteriophage T4 genome. Microbiol Mol Biol Rev. 2003;67(1):86–156.
33. Pope WH, Jacobs-Sera D. Annotation of bacteriophage genome sequences using DNA master: An overview. In: Clokie MR, Kropinski AM, Lavigne R; eds. *Bacteriophages: Methods and Protocols: Volume 3*. Valley Stream, NY: Humana Press; 2018: 217–229.
34. Rutherford K, Parkhill J, Crook J, et al. Artemis: Sequence visualization and annotation. Bioinformatics. 2000;16(10): 944–945.
35. Okonechnikov K, Golosova O, Fursov M, et al. Unipro UGENE: A unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–1167.
36. Kropinski AM, Borodovsky M, Carver TJ, et al. In silico identification of genes in bacteriophage DNA. In: Clokie MR, Kropinski AM; *Bacteriophages: Methods and Protocols*. Valley Stream, NY: Humana Press; 2009: 57–89.
37. Kropinski AM, Waddell T, Meng J, et al. The host-range, genomics and proteomics of *Escherichia coli* O157:H7 bacteriophage rV5. Virol J. 2013;10(1):76.
38. Cahill J, Young R. Phage lysis: Multiple genes for multiple barriers. Adv Virus Res. 2019;103:33–70.
39. Young R. Phage lysis: Three steps, three choices, one outcome. J Microbiol. 2014;52(3):243–258.
40. Kongari R, Rajaure M, Cahill J, et al. Phage spanins: Diversity, topological dynamics and gene convergence. BMC Bioinformatics. 2018;19(1):326.
41. Kelley DS, Lennon CW, Belfort M, et al. Mycobacteriophages as incubators for intein dissemination and evolution. MBio. 2016;7(5):e01537-16.
42. Edgell DR, Gibb EA, Belfort M. Mobile DNA elements in T4 and related phages. Virol J. 2010;7(1):290.
43. Lavigne R, Vandersteegen K. Group I introns in Staphylococcus bacteriophages. Future Virol. 2013;8(10):997–1005.
44. Xu S. Sequence-specific DNA nicking endonucleases. Biomol Concepts. 2015;6(4):253–267.
45. Kropinski AM, Arutyunov D, Foss M, et al. Genome and proteome of *Campylobacter jejuni* bacteriophage NCTC 12673. Appl Environ Microbiol. 2011;77(23):8265–8271.
46. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–D745.
47. Meydan S, Klepacki D, Mankin AS, et al. Identification of translation start sites in bacterial genomes. In: Labunskyy VM, ed. Ribosome Profiling. Methods in Molecular Biology, Volume 225. New York: Humana Press; N2021: 2–55.
48. Villegas A, Kropinski AM. An analysis of initiation codon utilization in the Domain Bacteria—Concerns about the quality of bacterial genome annotation. Microbiology. 2008;154(9):2559–2661.
49. Lazeroff M, Ryder G, Harris SL, et al. Phage commander, an application for rapid gene identification in bacteriophage genomes using multiple programs. Phage. 2021; 2(4):204–213.
50. Krishnamurthy SR, Wang D. Extensive conservation of prokaryotic ribosomal binding sites in known and novel picobirnaviruses. Virology. 2018;516:108–114.
51. Bailey TL, Johnson J, Grant CE, et al. The MEME suite. Nucleic Acids Res. 2015;43(W1):W39–W49.
52. Blum M, Chang HY, Chuguransky S, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2021;49(D1):D344–D354.
53. Lu S, Wang J, Chitsaz F, et al. CDD/SPARCLE: The conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265–D268.
54. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005;33:W244–W248.
55. Kall L, Krogh A, Sonnhammer EL. Advantages of combined transmembrane topology and signal peptide prediction—The Phobius web server. Nucleic Acids Res. 2007;35:W429–W432.
56. Sonhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6:175–182.
57. MacDougall A, Volynkin V, Saidi R, et al. UniRule: A unified rule resource for automatic annotation in the UniProt Knowledgebase. Bioinformatics. 2020;36:4643–4648.
58. Hulo C, de Castro E, Masson P, et al. ViralZone: A knowledge resource to understand virus diversity. Nucleic Acids Res. 2011;39(S1):576–582.
59. Tynecki P, Guzinski A, Kazimierczak J, et al. PhageAI—Bacteriophage life cycle recognition with machine learning and natural language processing. bioRxiv .2020 DOI 10.1101/2020.07.11.198606.
60. McNair K, Bailey BA, Edwards RA. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics. 2012;28(5):614–618.
61. Hockenberry AJ, Wilke CO. BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains. Peer J. 2021;9:e11396.
62. Fišarová L, Botka T, Du X, et al. Staphylococcus epidermidis phages transduce antimicrobial resistance plasmids and mobilize chromosomal islands. mSphere. 2021; 6(3):e00223-21.
63. Lekunberri I, Subirats J, Borrego CM, et al. Exploring the contribution of bacteriophages to antibiotic resistance. Environ Pollut. 2017;220:981–984.
64. Moon K, Jeon JH, Kang I, et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. Microbiome. 2020;8(1):75.
65. Yang Y, Xie X, Tang M, et al. Exploring the profile of antimicrobial resistance genes harboring by bacteriophage in chicken feces. Sci Total Environ. 2020;700:134446.
66. Brown-Jaque M, Calero-Cáceres W, Espinal P, et al. Antibiotic resistance genes in phage particles isolated from human faeces and induced from clinical bacterial isolates. Int J Antimicrob Agents. 2018;51(3):434–442.
67. Enault F, Briet A, Bouteille L, et al. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. ISME J. 2017;11(1):237–247.

68. Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 2019; 48:D517–D525.
69. Feldgarden M, Brover V, Gonzalez-Escalona N, et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):12728.
70. Nadiras C, Eveno E, Schwartz A, et al. A multivariate prediction model for Rho-dependent termination of transcription. Nucleic Acids Res. 2018;46(16):8245–8260.
71. McLean BW, Wiseman SL, Kropinski AM. Functional analysis of sigma-70 consensus promoters in Pseudomonas aeruginosa and Escherichia coli. Can J Microbiol. 1997;43(10):981–985.
72. Kozakai T, Izumi A, Horigome A, et al. Structure of a Core Promoter in Bifidobacterium longum NCC2705. J Bacteriol. 2020;202(7):e00540-19.
73. Lara-Gonzalez S, Dantas Machado AC, Rao S, et al. The RNA polymerase α subunit recognizes the DNA shape of the upstream promoter element. Biochemistry. 2020; 59(48):4523–4532.
74. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13): 3406–3415.
75. Dreher TW. Viral tRNAs and tRNA-like structures. Wiley Interdiscip. Rev RNA. 2010;1(3):402–414.
76. Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA genes in genomic sequences. In: Kollmar M, ed. Gene Prediction. Methods in Molecular Biology, vol 1962. New York: Humana Press; 2019: 1–14.
77. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–16.
78. Adriaenssens E, Brister JR. How to name and classify your phage: An informal guide. Viruses. 2017;9(4):70.
79. Russell DA, Hatfull GF. PhagesDB: The actinobacteriophage database. Bioinformatics 2016;33(5):784–786.
80. Hanauer DI, Graham MJ, SEA-PHAGES, et al. An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. Proc Natl Acad Sci. 2017;114(51): 13531–13536.
81. Walker PJ, Siddell SG, Lefkowitz EJ, et al. Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). Arch Virol. 2021;166(9):2633–2648.
82. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. Viruses. 2021;13(3): 506.
83. Moraru C, Varsani A, Kropinski AM. VIRIDIC—A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. Viruses. 2020;12(11):1268.
84. Johnson M, Zaretskaya I, Raytselis Y, et al. NCBI BLAST: A better web interface. Nucleic Acids Res. 2008; 36:W5–W9.
85. Clokie MR, Kropinski AM, Lavign R. *Bacteriophages: Methods and Protocols, Volume 3*. 1681. New York: Humana Press; 2018.
86. Russell DA. Sequencing, assembling, and finishing complete bacteriophage genomes. In: Clokie MR, Kropinski AM, Lavigne R; eds. *Bacteriophages: Methods and Protocols, Volume 3*. New York: Humana Press; 2018: 109–125.
87. Zhang X, Wang Y, Tong Y. Analyzing genome termini of bacteriophage through high-throughput sequencing. In: Clokie MR, Kropinski AM, Lavigne R; eds. *Bacteriophages: Methods and Protocols, Volume 3*. New York: Humana Press; 2018: 139–163.
88. Turner D, Sutton JM, Reynolds DM, et al. Visualization of phage genomic data: Comparative genomics and publication-quality diagrams. In: Clokie MR, Kropinski AM, Lavigne R; eds. *Bacteriophages: Methods and Protocols, Volume 3*. New York: Humana Press; 2018: 239–260.
89. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–2069.
90. Aziz RK, Bartels D, Best AA, et al. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics. 2008;9(1):75.
91. McNair K, Aziz RK, Pusch GD, et al. Phage genome annotation using the RAST pipeline. In: Clokie M, Kropinski A, Lavigne R; eds. *Bacteriophages: Methods in Molecular Biology*. New York: Humana Press; 2018: 231–238.
92. Davis JJ, Wattam AR, Aziz RK, et al. The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. Nucleic Acids Res. 2020;48(D1): D606–D612.
93. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34(D1):1037–1039.
94. Tanizawa Y, Fujisawa T, Arita M, et al. Generating publication-ready prokaryotic genome annotations with DFAST. In: Kollmar M, ed. Gene Prediction. Methods in Molecular Biology, vol 1962. New York: Humana Press; 2019: 215–226.
95. Ramsey J, Rasche H, Maughmer C, et al. Galaxy and Apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation. PLoS Comput Biol. 2020;16(11):e1008214.
96. Ecale Zhou CL, Malfatti S, Kimbrel J, et al. multiPhATE: Bioinformatics pipeline for functional annotation of phage isolates. Bioinformatics. 2019;35(21):4402–4404.
97. Burland TG. DNASTAR's Lasergene Sequence Analysis Software. In: Misener S, Krawetz SA; eds. *Bioinformatics: Methods and Protocols*. Totowa, NJ: Humana Press; 2000: 71–91.
98. Kearse M, Moir R, Wilson A, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–1649.
99. Joensen KG, Scheutz F, Lund O, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol. 2014;52(5):1501–1510.
100. Garg A, Gupta D. VirulentPred: A SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics. 2008;9(1):62.
101. Chakraborty A, Ghosh S, Chowdhary G, et al. DBETH: A database of bacterial exotoxins for human. Nucleic Acids Res. 2012;40(D1):D615–D620.
102. Wishart D, Arndt D, Pon A, et al. T3DB: The toxic exposome database. Nucleic Acids Res. 2015;43(D1): D928–D934.

103. Liu B, Zheng D, Jin Q, et al. VFDB 2019: A comparative pathogenomic platform with an interactive web interface. Nucleic Acids Res. 2019;47(D1):D687–D692.
104. Klucar L, Stano M, Hajduk M. phiSITE: Database of gene regulation in bacteriophages. Nucleic Acids Res. 2010;38: D366–D370.
105. Sampaio M, Rocha M, Oliveira H, et al. Predicting promoters in phage genomes using PhagePromoter. Bioinformatics. 2019;35(24):5301–5302.
106. Baerends RS, Smits WK, de Jong A, et al. Genome2D: A visualization tool for the rapid analysis of bacterial transcriptome data. Genome Biol. 2004;5(5):R37.
107. Coppens L, Lavigne R. SAPPHIRE: A neural network based classifier for σ70 promoter prediction in Pseudomonas. BMC Bioinformatics. 2020;21(1):415.
108. Lavigne R, Sun WD, Volckaert G. PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. Bioinformatics. 2004;20(5):629–635.
109. Bailey TL. STREME: Accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–2840.
110. Naville M, Ghuillot-Gaudeffroy A, Marchais A, et al. ARNold: A web tool for the prediction of Rho-independent transcription terminators. RNA Biol. 2011; 8(1):11–13.
111. Solovyev V, Salamov A. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In: Li RW; ed. *Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies*. Hauppauge, NY: Nova Science Publishers, Inc.; 2011: 61–78.
112. Feng CQ, Zhang ZY, Zhu XJ, et al. iTerm-PseKNC: A sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics. 2019;35(9):1469–1477.
113. Fan Y, Wang W, Zhu Q. iterb-PPse: Identification of transcriptional terminators in bacterial by incorporating nucleotide properties into PseKNC. PLoS One. 2020; 15(5):e0228479.
114. Stothard P. The sequence manipulation suite: JavaScript Programs for analyzing and formatting protein and DNA sequences. Biotechniques. 2000;28(6):1102–1104.
115. Rajaure M, Berry J, Kongari R, et al. Membrane fusion during phage lysis. Proc Natl Acad Sci USA. 2015; 112(17):5497–5502.
116. Perler FB. InBase, the intein database. Nucleic Acids Res. 2000;28(1):344–345.

Address correspondence to:
Evelien M. Adriaenssens, PhD
Quadram Institute Bioscience
Norwich Research Park
Norwich NR4 7UQ
United Kingdom

E-mail: evelien.adriaenssens@quadram.ac.uk