Phage Genome Annotation Using the RAST Pipeline

Katelyn McNair, Ramy Karam Aziz, Gordon D. Pusch, Ross Overbeek, Bas E. Dutilh, and Robert Edwards

Abstract

Phages are complex biomolecular machineries that have to survive in a bacterial world. Phage genomes show many adaptations to their lifestyle such as shorter genes, reduced capacity for redundant DNA sequences, and the inclusion of tRNAs in their genomes. In addition, phages are not free-living, they require a host for replication and survival. These unique adaptations provide challenges for the bioinformatics analysis of phage genomes. In particular, ORF calling, genome annotation, noncoding RNA (ncRNA) identification, and the identification of transposons and insertions are all complicated in phage genome analysis. We provide a road map through the phage genome annotation pipeline, and discuss the challenges and solutions for phage genome annotation as we have implemented in the rapid annotation using subsystems (RAST) pipeline.

Key words Phage, Genome annotation, RAST, Functional annotation, Gene predictions

1 The Steps of Phage Genome Annotation

The essential steps in annotating any genome, whether phage, bacterial, or eukaryotic, consist of identifying the features in the genome and assigning terms describing roles or functions to those features. Typical features that can be found in a phage genome include protein-encoding genes, noncoding RNA genes, insertion elements and transposons, direct and indirect repeats, origins of replication, and attachment or integration sites. Annotations are routinely only added to protein and RNA-encoding genes, labels are often provided for insertion elements or transposons. Specific for phages, they are fundamentally dependent on a cellular host to replicate, and the functions on its genome can only be completely understood in the context of the genome of the host. Thus, identification of prediction of the bacterial or archaeal host is an important part of phage annotation. Together, these features provide the core annotation of phages and this annotation provides the first steps to understanding the function of the phage as it interacts with

Martha R.J. Clokie et al. (eds.), *Bacteriophages: Methods and Protocols, Volume 3*, Methods in Molecular Biology, vol. 1681, https://doi.org/10.1007/978-1-4939-7343-9_17, © Springer Science+Business Media LLC 2018

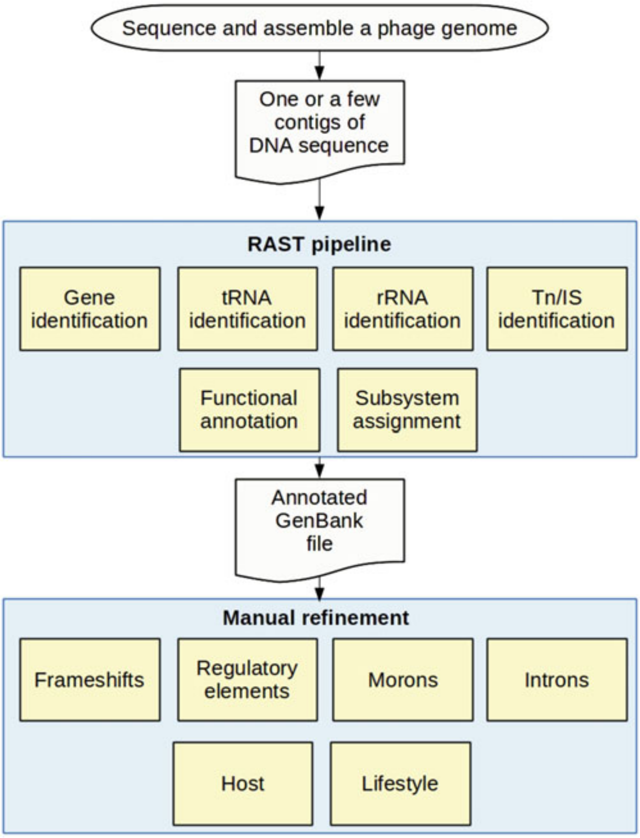

Fig. 1 Pipeline of phage genome annotation starting with DNA sequences and ending with an annotated genome

its host (Fig. 1). We discuss the approaches to identify and annotate each of these features below, and discuss how these annotations are performed in the Rapid Annotation Using Subsystems Technology approach (RAST) [1, 2].

Protein-encoding genes are the focus of most automated annotation systems, and more algorithms have been designed to handle these features than other features. Generally a protein-encoding gene can be identified as a long stretch of sequence in one reading frame that can be translated into protein sequence without including one of the three stop codons; these long stretches are called Open Reading Frames (ORFs). In gene calling, the stop codons are obvious because there is a choice of three codons to choose from and they are all stop codons (unless the phage encodes a suppressor tRNA which we do not discuss here). Most algorithms attempt to identify the longest nonoverlapping ORFs in a genome, based on the theory that the longer the open reading frame the less likely it is to occur by chance. There are many alternative gene-finding algorithms that have been developed over the last two decades, including CRITICA [3], GeneMark [4, 5], GISMO [6], Glimmer [7, 8], MetaGeneAnnotator [9], and Prodigal [10]. Most of the gene-finding algorithms find the same large genes because these are obvious and have high confidence. The algorithms may differ in the particular start sites that they identify; there may be multiple methionine (ATG) or valine (GTG) codons that could all be used as the start codon, and predicting exactly which start codon is the correct one for a given gene is difficult without a priori knowledge of the translation boundaries of the gene. In addition, the gene callers also differ in their ability to identify small protein-encoding genes. Short genes are statistically difficult to separate from the background noise of stretches of nucleotides that do not encode a stop codon, and often gene calling algorithms use an artificial cut off of (for example) 75 amino acids. It remains to be determined how many small proteins are encoded in phage genomes, and this is unlikely to be approached from a pure bioinformatics standpoint, as it will require biological validation of bioinformatics predictions or large-scale proteomic studies.

Most bacterial genomes are not thought to contain overlapping open reading frames, and these *shadow ORFs* are removed during the annotation step [10]. In viruses, including phages, however, there are several well-known examples of two different genes from the same stretch of DNA, such as the Rz/Rz1 system [11]. One study even suggests that new genes may be born via this process, providing evidence from the comparative genomics of *Rhabdoviridae* genomes [12]. These overlapping regions are generally not predicted using most bioinformatics approaches, as adding overlapping ORFs to gene prediction algorithms would include an enormous number of false positives to compensate for only a few false negatives. Therefore, most phage protein prediction schemes ignore overlapping proteins.

Following ORF identification, most bioinformatic gene prediction tools assign a confidence score to the ORFs using a model of what a gene is expected to look like, based on its nucleotide usage statistics. These statistics are specific for a species, and depend on properties like the codon usage and GC content of the genome. In bacterial genomes, the RAST pipeline starts by identifying highly conserved genes that are present in nearly every genome. The statistics from those genes are then used to build a genome-specific model for open reading frame identification that is applied to the rest of the genome. In phage genes, there are typically very few, if any, highly conserved genes, and never enough to build a reliable gene model. Therefore, most gene calling is performed by a generic model that is not trained on the specific genome being annotated but on the genomes of all phages. By default, the RAST pipeline uses Glimmer to identify the open reading frames, but options are available to use MetaGeneAnnotator [9], GeneMark [4], or Prodigal [10].

The functional annotation of protein-encoding phage genes is usually based on homology searches against existing phages. Historically, phage genes were named with a single letter starting at gpA, and either proceeding along the genome or assigning names based on the order in which the genes or their products were found. This resulted in several unrelated proteins from different phages all having the same names. For example both terminases and DNA replication initiation proteins have been annotated as gpA in different phage genomes available from GenBank. This confusion, amplified by the explosion of genome sequences in recent years, led to efforts to categorize phage proteins into either phage orthologous groups (POGs) [13] or subsystems [14] that have unified the annotation of many phage proteins. These common, descriptive, names provide a framework for comparing annotations among different phage genomes. The RAST system uses a combination of homology, chromosomal clustering, and subsystems to assign functions to proteins. First, proteins are annotated on the basis of homology to known proteins. If this initial search yields matches to proteins that are a component of a subsystem, RAST then tries to find other members of the subsystem that should be present in the same genome based on information from the previously annotated genomes. The advantage of this approach is that the RAST system can strengthen otherwise weak assertions of homology, based on predictions from subsystem annotations. Of note, the RAST tools allow the analysis of proteins in their chromosomal context, which sometimes helps determine the roles of proteins with unknown functions based on the functions of their chromosomal neighbors (e.g., protein subunits encoded by different genes, members of operons, or transporters of metabolites whose metabolizing enzymes are encoded on the same cluster). Phage genomes, like bacterial genomes, also order some of their genes, and this information can be leveraged to identify clusters of genes. For example, the small and large terminase (TerS and TerL) are frequently adjacent on the genome, and the identification of one leads to the identification of the other.

A major difficulty in the functional annotation of protein-encoding genes on phage genomes by homology searches is the fact that most proteins have no close homologs in the reference databases. Especially for novel phages, this results in the majority of encoded ORFs having no annotated function, or a hypothetical function at best. A possible solution includes homology-independent annotation, based on amino acid usage profiles of the proteins. One such approach, iVIREONS (https://vdm.sdsu.edu/ivireons/) uses machine learning to "learn" the characteristics of manually annotated phage proteins and then tests unknown proteins to see if they have similar characteristics [15].

Noncoding RNA (ncRNA) genes. Although Ribosomal RNAs have not yet been found in phage genomes, most pipelines, including RAST, look for them anyway as the pipelines have been developed for bacterial genome annotation and the computational cost of looking for rRNA genes is a minimal addition to the pipeline. Ribosomal RNA genes are highly conserved and are identified by extrinsic gene calling—using a database of known RNA genes to compare against. In contrast to rRNA genes that are recognized by homology, tRNA genes are recognized by intrinsic gene calling—using only features of the sequence. They are typically identified by computational tools built specifically to recognize the secondary structure of the tRNA molecule [16]. As with tRNAs, the function of other non-protein coding RNA genes also depends on the structure of the folded RNA molecule rather than the nucleotide sequence. Therefore, other noncoding RNA genes are also recognized by their conserved secondary structure rather than homology to existing sequences [17]. The RAST pipeline uses a manually curated database of ribosomal RNA genes to find them in a genome, and uses tRNAScan-SE [16] to identify tRNA genes. Many phages encode tRNA genes, and it has been proposed that these may supplement host-encoded tRNAs in translating phage proteins for anticodons that are insufficiently covered by the bacterial tRNAs [18]. These tRNA genes are also often used as phage integration sites in the host's genome (*attP*). Integration of the phage disrupts the host gene, and thus carrying complete, or near complete, tRNA genes allows the phage to reconstitute a tRNA into which it can integrate [19]. There has been little exploration of the role of ncRNA in phage lifestyle. Recent work with CRISPR/Cas systems have identified the presence of these systems in phage genomes [20] and metagenomes [21], and it is thought that they are being used to attack other phages that may be infecting the same host.

Insertion elements and transposons are currently identified by annotations of protein-encoding genes. Transposases (Tn) are readily identified as protein-encoding genes, and the similarity between members of the transposase family, and with other recombinases, is high enough that they usually receive accurate annotation. However, the repeats flanking the insertion sequence or transposon are not typically automatically annotated. There are boutique databases of these problematic mobile elements [22, 23], but often the classification of insertion (IS) elements is dependent on one or a few residues. Typically automatic annotation systems identify the Tn or IS elements but cannot identify the fine details responsible for the accurate categorization of these elements. More work is required to accurately denote the ends of these mobile elements in automatic phage annotation systems. Direct and indirect repeats are usually used to identify the ends of insertion elements and transposons [22], and to predict the ends of prophages that have been found in bacterial genomes [13]. Standard informatics approaches can easily identify repeats longer than approximately 14 nucleotides in a phage genome. Below that length, repeats are found too frequently to ascertain whether they are indeed the correct flanking repeats, or randomly occurring repeated sequence elements. A few websites can be used to identify repeats in DNA sequences (e.g., [24, 25]).

Phage attachment sites are impossible to detect de novo if only the phage is known, but if the phage and the host genome sequences are known, they are trivial to find. The phage carries the attachment site P (*attP*) that has sequence homology to the bacterial attachment site B (*attB*). Integration is initiated by recombination between *attP* and *attB*, resulting in *attL* and *attR* sites that flank the nascent prophage.

Accurately Annotating Phage Metadata. Annotating genomic metadata is a general challenge to genomics and metagenomics. With bacteriophages, this issue is even more problematic, given the lack of systematic nomenclature for viruses (as opposed to the binomial system used for cellular organisms, *see* Chapter 15 of this book). Some attempts were made to suggest systematic nomenclature for viruses similar to those used for plasmids [26], but they are not widely applied or enforced. In addition to accurate taxonomic descriptions of viruses, including metadata associated with the virus (e.g., its morphology, actual host, host range, and lifestyle) is equally important. These make comparative genomics studies possible, enable predictive tools such as those that identify the host of unknown phages [27], or predict the lifestyle of new phages [14] and improve metagenomic/microbiomic annotations. Other important types of metadata can be computed from the genomic information, e.g., a genome's length, %G+C, and codon usage [28]. These too have quite powerful applications in comparative genomics, prophage finding, and metagenomics. For example, information content of phage genomes has improved prophage finding [29] and is proposed to improve metagenomic analysis [30]. As with gene annotation, metadata annotation needs to use a controlled vocabulary (which has to be consistent but not necessarily rigid or hierarchical). Spelling inconsistencies (e.g., firmicutes vs. Firmicutes vs. gram-positive bacteria) or terminology inconsistencies (e.g., temperate vs. lysogenic lifestyles) are all obstacles against computational analysis and data propagation.

To summarize, phage annotation involves the identification and functional description of several types of features, including protein-encoding genes, RNA genes, insertion elements and transposons, repeats, and attachment sites. Moreover, phage–host associations are an important part of understanding phage biology that can be predicted using a range of computational tools [27]. The RAST pipeline provides an automated approach to phage genome annotation. The pipeline currently uses bacterial ORF-finding algorithms to identify the proteins in the genome, and a combination of homology-based and subsystems-based approaches to decorate those proteins with their functional annotation. RNA genes are detected by a combination of extrinsic and intrinsic gene calling methods. There remain several hurdles to accurate phage genome annotation, especially the assignment of functions to unknown proteins, the identification of small proteins in the genome, and the correct and unambiguous identification of insertion elements and transposons. The combinations of bioinformatics advances and a better understanding of phage biology will help to improve phage genome annotation, making this field a fertile area for further exploration.

Acknowledgments

This work was supported by grants from the National Science Foundation MCB-1330800 and DUE-1323809 to RAE. BED was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004.

References

1. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75
2. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason Iii JA, Stevens R, Vonstein V, Wattam AR, Xia F (2015) RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. Sci Rep 5:8365
3. Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol 16:512–524
4. Borodovsky M, Mclninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res 23:3554–3562
5. Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115
6. Krause L, McHardy AC, Pühler A, Stoye J, Meyer F (2007) GISMO - Gene identification using a support vector machine for ORF classification. Nucleic Acids Res 35:540–549
7. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–4641
8. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 40:e9–e9
9. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 15:387–396
10. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119
11. Summer EJ, Berry J, Tran TAT, Niu L, Struck DK, Young R (2007) Rz/Rz1 lysis gene equivalents in phages of Gram-negative hosts. J Mol Biol 373:1098–1112
12. Walker PJ, Firth C, Widen SG, Blasdell KR, Guzman H, Wood TG, Paradkar PN, Holmes EC, Tesh RB, Vasilakis N (2015) Evolution of genome size and complexity in the *Rhabdoviridae*. PLoS Pathog 11:e1004664

13. Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV (2013) Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. J Bacteriol 195:941–950
14. McNair K, Bailey BA, Edwards RA (2012) PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28:614–618
15. Seguritan V, Alves N, Arnoult M, Raymond A, Lorimer D, Burgin AB, Salamon P, Segall AM (2012) Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput Biol 8:e1002657
16. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964
17. Nawrocki EP (2014) Annotating functional RNAs in genomes using Infernal. Methods Mol Biol 1097:163–197
18. Bailly-Bechet M, Vergassola M, Rocha E (2007) Causes for the intriguing presence of tRNAs in phages. Genome Res 17:1486–1495
19. Williams KP (2002) Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res 30:866–875
20. Seed KD, Lazinski DW, Calderwood SB, Camilli A (2013) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494:489–491
21. Cassman N, Prieto-Davó A, Walsh K, Silva GGZ, Angly F, Akhter S, Barott K, Busch J, McDole T, Haggerty JM, Willner D, Alarcón G, Ulloa O, DeLong EF, Dutilh BE, Rohwer F, Dinsdale EA (2012) Oxygen minimum zones harbour novel viral communities with low diversity. Environ Microbiol 14:3043–3065
22. Aziz RK, Breitbart M, Edwards RA (2010) Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res 38:4207–4217
23. Riadi G, Medina-Moenne C, Holmes DS (2012) TnpPred: a web service for the robust prediction of prokaryotic transposases. Comp Funct Genomics 2012:678761
24. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580
25. Volfovsky N, Haas BJ, Salzberg SL (2001) A clustering method for repeat analysis in DNA sequences. Genome Biol 2:RESEARCH0027
26. Kropinski AM, Prangishvili D, Lavigne R (2009) Position paper: the creation of a rational scheme for the nomenclature of viruses of Bacteria and Archaea. Environ Microbiol 11:2775–2777
27. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE (2016) Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol Rev 40:58–72
28. Aziz RK, Dwivedi B, Akhter S, Breitbart M, Edwards RA (2015) Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. Front Microbiol 6:381
29. Akhter S, Aziz RK, Edwards RA (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res 40:e126–e126
30. Akhter S, Bailey BA, Salamon P, Aziz RK, Edwards RA (2013) Applying Shannon's information theory to bacterial and phage genomes and metagenomes. Sci Rep 3:1033