BlueFeather, the singleton that wasn't: Shared gene content analysis supports expansion of *Arthrobacter* phage cluster FE

Stephanie Demo¶, Andrew Kapinos¶, Aaron Bernardino, Kristina Guardino, Blake Hobbs, Kimberly Hoh, Edward Lee, Iphen Vuong, Krisanavane Reddi, Amanda C. Freise, Jordan Moberg Parker*

¶*Denotes equal contributions*

Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Keywords: bacteriophage, *Arthrobacter*, clustering, virology, microbiology, comparative genomics

*Corresponding Author:
Jordan Moberg Parker, jmobergparker @ ucla.edu

Brief Running Title

Arthrobacter phage BlueFeather

Abstract

Bacteriophages (phages) exhibit high genetic diversity, and the mosaic nature of the shared genetic pool makes quantifying phage relatedness a shifting target. Early parameters for clustering of related *Mycobacteria* and *Arthrobacter* phage genomes relied on nucleotide identity thresholds but, more recently, clustering of *Gordonia* and *Microbacterium* phages has been performed according to shared gene content. Singleton phages lack the nucleotide identity and/or shared gene content required for clustering newly sequenced genomes with known phages. Whole genome metrics of novel *Arthrobacter* phage BlueFeather, originally designated a putative singleton, showed low nucleotide identity but high amino acid and gene content similarity with *Arthrobacter* phages originally assigned to Clusters FE and FI. Gene content similarity revealed that BlueFeather shared genes with these phages in excess of the parameter for clustering *Gordonia* and *Microbacterium* phages. Single gene analyses revealed evidence of horizontal gene transfer between BlueFeather and phages in unique clusters that infect a variety of bacterial hosts. Our findings highlight the advantage of using shared gene content to study seemingly genetically isolated phages and have resulted in the reclustering of BlueFeather, a putative singleton, as well as former Cluster FI phages, into a newly expanded Cluster FE.

Introduction

Bacteriophages are ubiquitous biological entities with an estimated 10^{31} phage particles on Earth. Assuming an average length of 200 nm, they would extend 200 million light years if stacked head-to-tail (Wiles 2014). Phages are found in all ecosystems in which bacteria exist and function as drivers of bacterial evolution (Keen

2015). They exhibit horizontal gene transfer (HGT) with each other and with bacteria, resulting in the diverse and mosaic nature of phage genomes (Pedulla et al. 2003). Despite their incredible prevalence in the environment, phages remain largely understudied (Miller-Ensminger et al. 2018).

Previous research on mycobacteriophages concluded that phages may exhibit a continuum of diversity, wherein genes are constantly being shuffled amongst the phage population, resulting in shared genes and sequences between different clusters (Pope et al. 2015). The immense and ever-expanding diversity of phage genomes has historically been categorized in terms of nucleotide sequence conservation, with a minimum 50% nucleotide identity and 50% span length to at least one phage in a cluster to warrant membership (Hatfull et al. 2010; Klyczek et al. 2017). A mass scale study on *Gordonia* phages also identified a spectrum of genetic diversity, as clusters did not have clear boundaries (Pope et al. 2017). Numerous phages lacked the requirement of 50% nucleotide identity but shared many genes, suggesting a relatedness not captured by nucleotide comparisons alone. This relatedness was confirmed with a gene content network phylogeny and subsequently, the cluster assignment parameter for *Gordonia* phages (Pope et al. 2017), and later for *Microbacterium* phages (Jacobs-Sera et al. 2020), was adjusted to 35% shared gene content with at least one phage in a cluster. Mycobacteriophages, as well as *Gordonia* and *Microbacterium* phages, exhibited this spectrum; however, the extent of diversity varies depending on the current known phage population, which in turn affects how clustering is carried out. *Arthrobacter* phages were previously found to exchange genes more slowly than *Gordonia* phages, and the 50% nucleotide clustering parameter was considered sufficient at the time (Pope

et al. 2017). Further studies on *Arthrobacter* phages found these phages to be genetically isolated with highly variable gene content for phages that can infect a range of host species. With this great diversity, nucleotide identity was used to separate *Arthrobacter* phages into 10 distinct clusters and 2 singletons (Klyczek et al. 2017), and this parameter has been considered sufficient to categorize the limited number of *Arthrobacter* phages until recently.

Singleton phages can serve as the seeds to start new clusters or be extremely distinct, as they lack the nucleotide identity and/or shared genes required for clustering with known phages. In this study, the genome of novel *Arthrobacter* phage BlueFeather was examined for nucleotide and amino acid identity with other known phages. BlueFeather lacked sufficient nucleotide conservation for clustering according to nucleotide-based parameters, and was thus designated a putative singleton. Phage BlueFeather did, however, have notable amino acid conservation and shared gene content with other *Arthrobacter* phages previously assigned to Clusters FE and FI, suggesting it may not be as isolated as its putative singleton status implied. The outcomes of this research on phage BlueFeather provided evidence for the reclustering of phage BlueFeather, as well as phages formerly assigned to Cluster FI, into a newly expanded Cluster FE.

Materials and methods

Sample collection and direct isolation

Soil was collected from Los Angeles, CA in a residential area located at 34.05638889° N, 118.445010000° W. Direct isolation of phages was performed by shaking a soil sample and 2X PYCa broth (Yeast Extract 1 g/L, Peptone 15 g/L, 4.5mM $CaCl_2$, Dextrose 0.1%) in conical tubes at 250 RPM at 25°C for 1.5 hours. After incubation, the solution was filtered through a 0.22 µm syringe and spotted onto *Arthrobacter globiformis* B-2979 (*A. globiformis*). Plaque purifications were performed and a high titer lysate was filter-sterilized to be used in subsequent characterization experiments.

Transmission electron microscopy

Transmission electron microscopy (TEM) was performed on BlueFeather lysate. The sample was placed onto a carbon-coated electron microscope grid and stained with 1% uranyl acetate. Phage particles were visualized using the CM120 Instrument (Philips, Amsterdam, Netherlands) and micrographs were captured. Phage head and tail lengths were measured using ImageJ (Schneider, Rasband, and Eliceiri 2012).

Genome sequencing and assembly

Viral DNA was isolated with the Wizard® DNA Clean-Up System (cat # A7280, Promega, WI, USA). Sequencing libraries were constructed with the NEBNext® Ultra™ II DNA Library Prep kit (New England Biolabs, MA, USA), and sequenced by Illumina-MiSeq at the Pittsburgh Bacteriophage Institute to an approximate shotgun coverage of 3538x. Genome assembly and finishing were performed as previously described (Russell 2018).

Gene annotation

Genomes were annotated as described previously (Pope and Jacobs-Sera 2018) using DNA Master (http://cobamide2.bio.pitt.edu/) and PECAAN (https://pecaan.kbrinsgd.org/) for auto-annotation. GLIMMER (Delcher et al. 1999) and GeneMark (Besemer and Borodovsky 2005) were used to predict protein-coding regions along with their start and stop sites. Manual annotation was performed using Phamerator (Cresawn et al. 2011), Starterator (*SEA-PHAGES/Starterator* [2016] 2020), and host-trained and self-trained GeneMark coding potential maps to support or refute auto-annotation predictions (Besemer and Borodovsky 2005). Gene functions were determined using PhagesDB BLAST (https://phagesdb.org/blastp/), NCBI BLAST (Altschul et al. 1990), HHpred (Soding, Biegert, and Lupas 2005) and CDD (Marchler-Bauer et al. 2014). The presence of transmembrane proteins was determined using TMHMM (Krogh et al. 2001) and TOPCONS (Tsirigos et al. 2015). The annotated complete genome was deposited to GenBank under the accession number MT024867.

Gene content comparisons

Phage genomes used in this study are available from phagesdb.org (Russell and Hatfull 2017). Gepard was used to perform sequence analysis to identify regions of homology between nucleotide sequences or amino acid sequences of different phages (Krumsiek, Arnold, and Rattei 2007). Concatenated whole genome nucleotide and whole amino acid sequences were used to create dot plots with word sizes of 15 and 5, respectively.

SplitsTree was used to generate a network phylogeny in order to reveal the genetic distance between *Arthrobacter* phages (Huson 1998). BlueFeather and up to 10

representative phages from each *Arthrobacter* cluster were selected from the Actino_Draft database (version 366) for comparison.

The gene content calculator on PhagesDB (https://phagesdb.org/genecontent/) was used to calculate Gene Content Similarity (GCS), the percentage of shared genes in phams (groups of genes with related sequences), between BlueFeather, Cluster FE, and former Cluster FI phages (Cresawn et al. 2011). Gene Content Dissimilarity (GCD) and maximum GCD gap (MaxGCDGap) were calculated using scripts described previously (Pope et al. 2017). Heatmaps and scatter plots were created using Prism 8.0.0 (GraphPad Software, San Diego, California, USA) and were used for quantitative analysis and visualization of GCS and GCD values.

PhagesDB Pham View was used to gather information about phages with genes in the same phams as BlueFeather's (Russell and Hatfull 2017). PECAAN was used to obtain the nucleotide sequences for each BlueFeather gene (https://discover.kbrinsgd.org). The BiologicsCorp online GC content calculator was used for each gene in the genome (https://www.biologicscorp.com/tools/GCContent/).

Results

BlueFeather is a siphovirus with a short genome

Phage BlueFeather was isolated from a soil sample via direct isolation on *A. globiformis* B-2979 at 25°C and had a bullseye plaque morphology 5 mm in diameter (Fig 1a). Transmission electron microscopy (TEM) at 67,000X magnification showed an average phage capsid diameter and tail length of 48 ± 8 nm and 156 ± 53 nm, respectively (Fig 1b). The long, flexible, non-contractile tail suggested BlueFeather's classification as a *Siphoviridae* (Yuan and Gao 2017).

BlueFeather's genome had a length of 16,302 bp, 64.30% GC content, and genome ends with 15 base 3' sticky overhangs (CCACGGTTCCCGTCC). Phages that infect *Arthrobacter* hosts have genome lengths that range from 15,319 bp (Toulouse) to 70,265 bp (PrincessTrina) (Klyczek et al. 2017). The average *Arthobacter* phage genome length (as of May 2020) was 46,968 bp with a standard deviation of 20,619 bp and a median length of 53,859 bp, suggesting that most *Arthrobacter* phages have genomes notably larger than that of BlueFeather. BlueFeather's genome contained 25 manually annotated genes; 18 were of known function, 6 were orphams–meaning they have not been identified in any other known phage–and 1 was a reverse gene (Fig 2). The left arm of the genome had highly conserved genes amongst siphoviral *Arthrobacter* phages, such as those encoding terminase, portal protein, head-to-tail adapter, and tail proteins (Klyczek et al. 2017). Tail tube and sheath genes were absent, confirming the classification of BlueFeather as a siphovirus. Genes characteristic of the lytic life cycle, such as lysin A and holin, were identified; however, there were no genes that would indicate BlueFeather's ability to undergo a lysogenic life cycle, suggesting that BlueFeather is not a temperate phage (Shi et al. 2012).

Dot plot comparisons revealed synonymous substitutions in BlueFeather's genome

Phage BlueFeather was originally classified as a singleton on PhagesDB due to low nucleotide identity with other known phages. Nucleotide and amino acid dot plots were created to qualitatively compare BlueFeather to the most similar *Arthrobacter* phages, including those in Cluster FE (Corgi, Idaho, Noely) and former Cluster FI (Whytu, Yavru), as identified by BLASTn. Due to the limited number of sequenced

Arthrobacter phages, many of the clusters have few members. Of the 28 *Arthrobacter* clusters on PhagesDB (as of May 2020), 17 clusters have between 2-4 phages (including the former Cluster FI). As expected, phages originally assigned to the same cluster had alignments indicating large regions of nucleotide similarity (Grose and Casjens 2014), while comparison of BlueFeather's genome to phages originally assigned to Clusters FE and FI revealed no homologous sequences (Fig 3a). Unexpectedly, dot plot analysis of concatenated amino acid sequences with a word size of 5 revealed numerous regions of amino acid sequence similarity between these phages (Fig 3b). This reflects, at present, perhaps one of the clearest examples in which a group of phages lack nucleotide identity while sharing considerable amino acid identity.

Gene similarity demonstrates a close relationship between BlueFeather and phages of Cluster FE and former Cluster FI

Gene Content Similarity (GCS) is a key metric in quantifying phage genetic relationships and is calculated by averaging the number of shared genes between two phages (Mavrich and Hatfull 2017). GCS was calculated for BlueFeather, Cluster FE phages, and phages originally assigned to Cluster FI. BlueFeather shared over 35% of genes with most Cluster FE phages, and over 55% of genes with former Cluster FI phages. Over 35% of genes were shared in each pairwise comparison performed, with the exception of phage Idaho (34.18% shared gene content with BlueFeather) (Fig 4a). Given that BlueFeather was originally determined to be a singleton, it was surprising to find GCS greater than the recently adopted threshold of 35% for clustering other phage populations (Jacobs-Sera et al. 2020; Pope et al. 2017). Gene Content Dissimilarity (GCD) is the opposite of GCS and was used to calculate the maximum GCD gap

(MaxGCDGap), a metric that represents the degree of isolation between a phage and a selected phage population (Pope et al. 2017). GCD was calculated for BlueFeather and all *Arthrobacter* phages. There was a MaxGCDGap of 44.45% between BlueFeather and Whytu, indicating a relatively high degree of separation between BlueFeather and the rest of the *Arthrobacter* phage population (Fig 4b). *Arthrobacter* phages which exhibited pairwise GCD values with BlueFeather of less than 1 were found in Clusters AN, AU, AM, AZ, AV, AL, FE, AO, FH, FF, and former Cluster FI, indicating shared gene content. GCD was then calculated for BlueFeather and all known phages in the PhagesDB Actino_draft database (Fig 4c). Similar to the *Arthrobacter* GCD plot, phages assigned to Cluster FE and former Cluster FI were the least dissimilar to BlueFeather. It is notable that in this comparison, there were 63 additional phages ranging from 0.959 to 0.975 GCD, meaning BlueFeather shares a low number of genes with many non-*Arthrobacter* phages. Non-*Arthrobacter* phages which exhibited pairwise GCD values with BlueFeather of less than 1 were found in *Microbacterium* phage Cluster EE, *Mycobacterium* phage Clusters N, I, P, and the singleton IdentityCrisis, as well as *Gordonia* phage Clusters DT, CW and the singleton GMA4.

To compare the relationships between the *Arthrobacter* phage population as whole and the phages assigned to Cluster FE, former Cluster FI, and BlueFeather, a SplitsTree network phylogeny of the phams from each *Arthobacter* phage cluster was generated to examine the genetic distance between the phages. As expected, BlueFeather was shown to be more genetically similar to phages originally assigned to Clusters FE and FI than to any other *Arthrobacter* phage clusters (Fig 5). BlueFeather demonstrated a closer pham similarity to former Cluster FI phages Whytu and Yavru

than to Cluster FE phages Idaho, Noely and Corgi; however, these phages altogether formed a distinct branch from the rest of the phages sampled and together comprise the newly expanded Cluster FE.

BlueFeather genome exhibits evidence of horizontal gene transfer

Given that BlueFeather shares genes with phages infecting distinct hosts, we investigated its genome for potential evidence of horizontal gene transfer (HGT). A whole genome heatmap was created using common metrics for evidence of HGT for each gene in the genome. As of March 2020, 4 genes in BlueFeather were considered to have the most convincing evidence for HGT based on GC content and prevalence in phages that infect unique bacterial hosts: genes 2, 15, 19, and 24 (Fig 6).

Typically, viral genes have about the same (Bohlin and Pettersson 2019) or slightly lower GC content (de Melo et al. 2019) compared to their bacterial hosts, suggesting that genes with higher GC content may have been horizontally transferred. BlueFeather had an overall average GC content of 64.30% and *Arthrobacter globiformis* mrc11 was found to have an overall GC content of 65.9% (Sahoo et al. 2019). BlueFeather gene-specific average GC contents ranged from 59.30% to 70.30%, and genes with maximum average GC contents were considered for HGT. This included genes 15 and 24 with GC contents of 70.3% and 70.1%, respectively.

It is increasingly understood that phages infecting different hosts may share considerable gene content through processes such as HGT (Pope et al. 2017). For each gene in the BlueFeather genome, we calculated the number of unique isolation hosts for phages possessing a pham found in BlueFeather. Gene 2 belongs to a pham with

member genes found in phages that infect *Gordonia malaquae* BEN700 and *Arthrobacter sp.* ATCC 21022. Gene 15 belongs to a pham with member genes found in phages that infect *A. globiformis* B-2979, *A. sp*. ATCC 21022, *Mycobacterium smegmatis* mc²155, *G. malaquae* BEN700, and *Gordonia rubripertincta* NRRL B-16540. Gene 19 was the only reverse gene in the BlueFeather genome, and this gene was only found in BlueFeather and in phages infecting *Microbacterium foliorum* NRRL B-24224 SEA and *Microbacterium paraoxydans* NWU1.

Discussion

Our research was focused on the genomic and evolutionary relationships between the novel *Arthrobacter* phage BlueFeather and other known phages, particularly those originally assigned to Clusters FE and FI. Previous studies have shown that new clusters can be formed when novel phages are found to be similar to former singletons, as demonstrated by the formation of Cluster AS from *Arthrobacter* singleton Galaxy (Klyczek et al. 2017). BlueFeather, originally designated as a putative singleton phage, exhibits over 35% GCS with nearly all phages originally in Cluster FE and over 55% GCS with those formerly assigned to Cluster FI. The conservation of amino acids, rather than nucleotides, suggests a history of purifying selection via many synonymous mutations in which deleterious mutations were filtered out (Ngandu et al. 2008). Moreover, this presents some of the most clear evidence to date of highly conserved amino acid sequences despite the absence of significant nucleotide conservation amongst related phages.

There is a high degree of synteny between these phages as well. While BlueFeather's original designation as a singleton would imply low genomic relatedness to other phages (Pope et al. 2015), gene similarity between BlueFeather and phages from Cluster FE and former Cluster FI–in excess of 35%–indicates conservation of gene functions and genome architecture despite extensive divergence of nucleotide identity. While *Arthrobacter* phages have been clustered according to nucleotide identity in the past (Klyczek et al. 2017), this study on BlueFeather, Cluster FE, and the former Cluster FI highlights the importance of continually reevaluating clustering parameters, particularly when different parameters may result in different cluster assignments. Moreover, BlueFeather has the smallest genome of all *Arthrobacter* singletons, and it is possible that clustering parameters may also need to take genome size into account. Given that GCS reflects the number of genes shared as a proportion of the total number of genes for each phage, the same number of shared genes would yield higher GCS in comparisons between smaller genomes.

Gene content dissimilarity demonstrated that BlueFeather has a MaxGCDGap of 44.45% with phage Whytu, which was originally assigned to Cluster FI. BlueFeather was found to be least dissimilar with Cluster FE and former Cluster FI phages; this was supported by a network phylogeny of representative *Arthrobacter* phages that indicated great diversity between clusters, but revealed that phage BlueFeather forms a distinct branch with Cluster FE and former Cluster FI phages. Additionally, many phages were found to share between 0-10% GCS with BlueFeather. While this is too low to warrant a significant phylogenetic relationship, it reinforced the observed continuum of diversity in phage populations. Previous research found *Arthrobacter* phage clusters to be very

discrete (Klyczek et al. 2017). Even so, this low yet seemingly widespread display of shared genes, as well as BlueFeather's unexpected relationships with Cluster FE and former Cluster FI phages, provides new insight into the genetic landscape of *Arthrobacter* phages. Few phages were previously assigned to Cluster FE and the former Cluster FI, representing only 5 of the 306 sequenced and manually annotated *Arthrobacter* phages (as of May 2020). On the other hand, there are 1,906 sequenced *Mycobacterium* phages (as of May 2020), which has allowed for a more thorough investigation of the mycobacteriophage continuum of diversity. As more *Arthrobacter* phages are sequenced, we expect to observe similar trends in these host-dependent genetic landscapes.

Unlike singleton phages that are replete with orphams (Klyczek et al. 2017; Pope et al. 2015), the BlueFeather genome, originally designated as a putative singleton, is composed predominantly of genes with known functions that have been assigned to phams. BlueFeather has less than half as many genes as current *Arthrobacter* singletons and contains highly conserved genes required for viral mechanisms. These vital functional genes have been more thoroughly studied and as a result, are more likely to be found in phams with predicted functions (Cresawn et al. 2011). Additionally, given that pham assignments are performed on the basis of amino acid identity, it is unsurprising that many of the phams containing these vital functional genes are shared amongst BlueFeather, Cluster FE, and former Cluster FI phages, despite the lack of significant nucleotide identity in gene encoding sequences.

Markers of horizontal gene transfer (HGT) included unexpectedly high GC content, as well as multiple bacterial hosts on which phages sharing genes with

BlueFeather were isolated (Pope et al. 2015). BlueFeather shared phams with a multitude of non-*Arthrobacter* phages from various clusters, which allowed us to identify multiple regions as having evidence for HGT. These potential HGT events serve to magnify phage diversity and promote the phenomenon of genetic mosaicism. BlueFeather serves as yet another example of the highly intricate mosaic relationships which exist among phages, a feature of the genetic landscape which makes phage taxonomy an increasingly difficult task.

In sum, this research has led to the reclustering of BlueFeather and phages formerly assigned to Cluster FI into a newly expanded Cluster FE. Recent observations in which there appear to be limited nucleotide conservation but high shared gene content, as observed in this newly expanded cluster, support the notion that clustering methods should be continually reevaluated and optimized as more phages are sequenced (Pope et al. 2017). This study thus provides valuable insight into the continuum of diversity amongst *Arthrobacter* phages, while also supporting a 35% shared gene content clustering parameter as was previously adopted for *Gordonia* and *Microbacterium* phages (Pope et al. 2017; Jacobs-Sera et al. 2020). Further investigation into novel phages is essential to understand the complex phage landscape. As more *Arthrobacter* phages are discovered, it is likely that we will discover many more phages like BlueFeather which belong to clusters whose close relationships become apparent only through the lens of shared gene content.

Acknowledgments

We thank Rebecca A. Garlena and Daniel A. Russell at the Pittsburgh Bacteriophage Institute for genome sequence and assembly; and Travis Mavrich,

Welkin Pope, Debbie Jacobs-Sera, and Graham Hatfull with the HHMI Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program for programmatic support. The authors acknowledge the use of instruments at the Electron Imaging Center for NanoMachines supported by NIH (1S10RR23057 to ZHZ) and CNSI at UCLA.

Author contributions

A.B., B.H., E.L., I.V., K.G., K.H., and S.D. performed experiments and drafted the paper; S.D., A.K., J.M.P. and A.F. revised the paper; and J.M.P. and A.F. supervised the research. This research was funded in part by the Microbiology, Immunology and Molecular Genetics Department, and the Dean of Life Sciences Division at UCLA.

Author disclosure statement

The authors declare that there is no conflict of interest regarding the publication of this article.

Figures

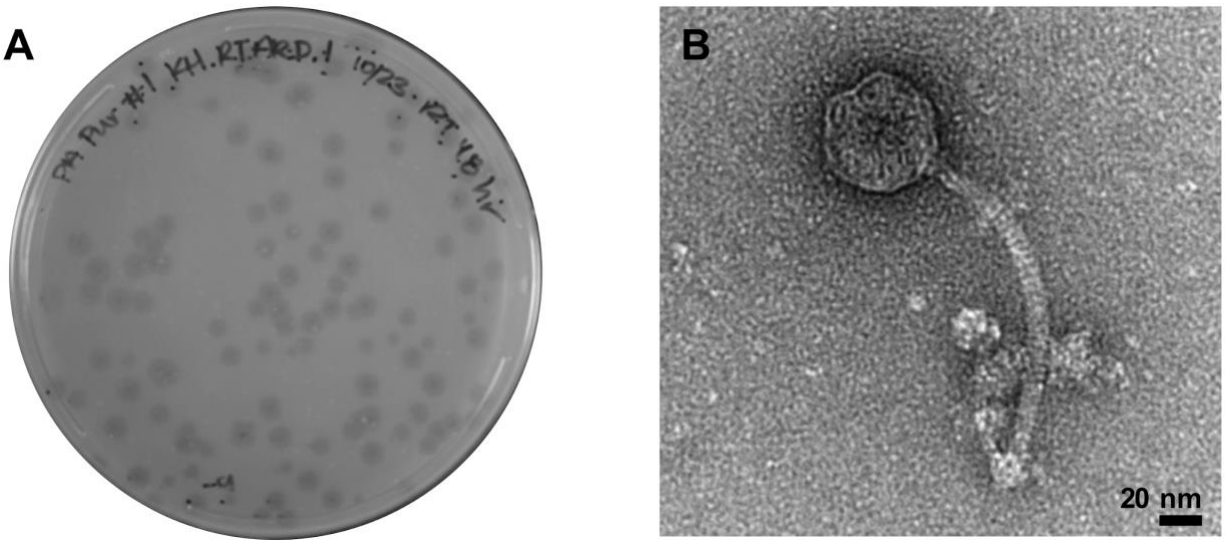


Fig 1. BlueFeather is a Siphovirus. (A) A picked plaque was suspended in phage buffer, then tested in a plaque assay for purification. Plaque morphology was consistent with bullseye plaques around 5 mm in diameter. **(B)** TEM image of BlueFeather at 67,000X magnification. The capsid was estimated to be 48 ± 8 nm and the tail 156 ± 53 nm.

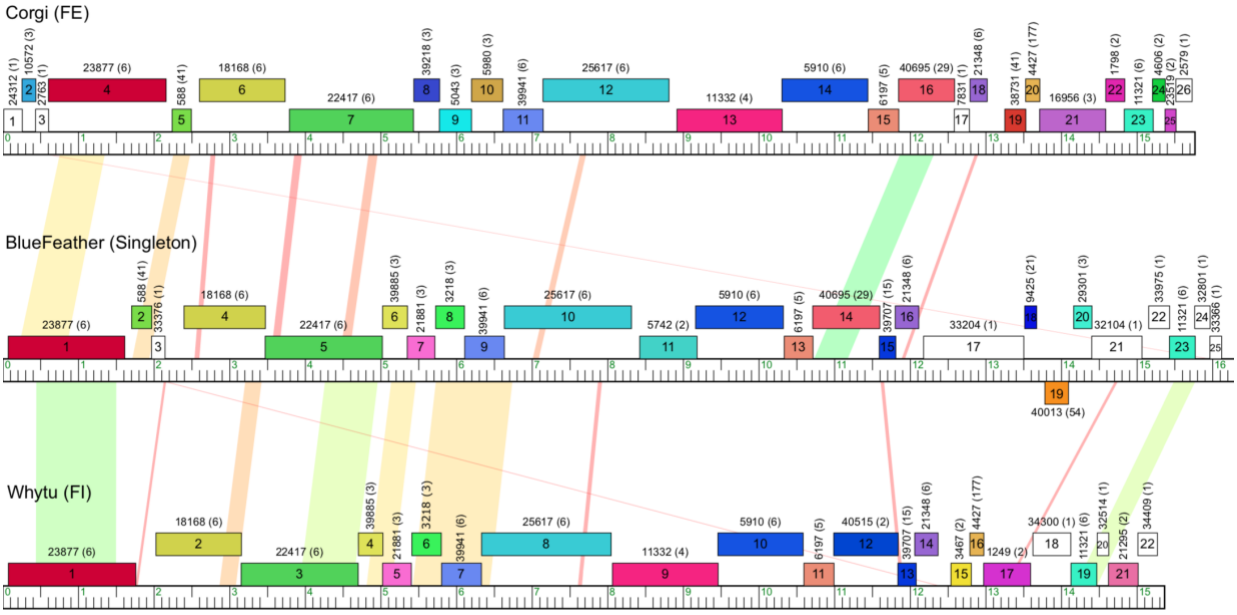


Fig 2. BlueFeather genome shares little nucleotide similarity but many phams with Cluster FE and former Cluster FI. The BlueFeather genome is linear with a relatively small length of 16 kbp. Of the 25 identified ORFs, 18 were of known function, 6 were orphams and 1 was a reverse gene. BlueFeather had little BLASTn homology to its most similar phages, as indicated by the limited orange and yellow shading.

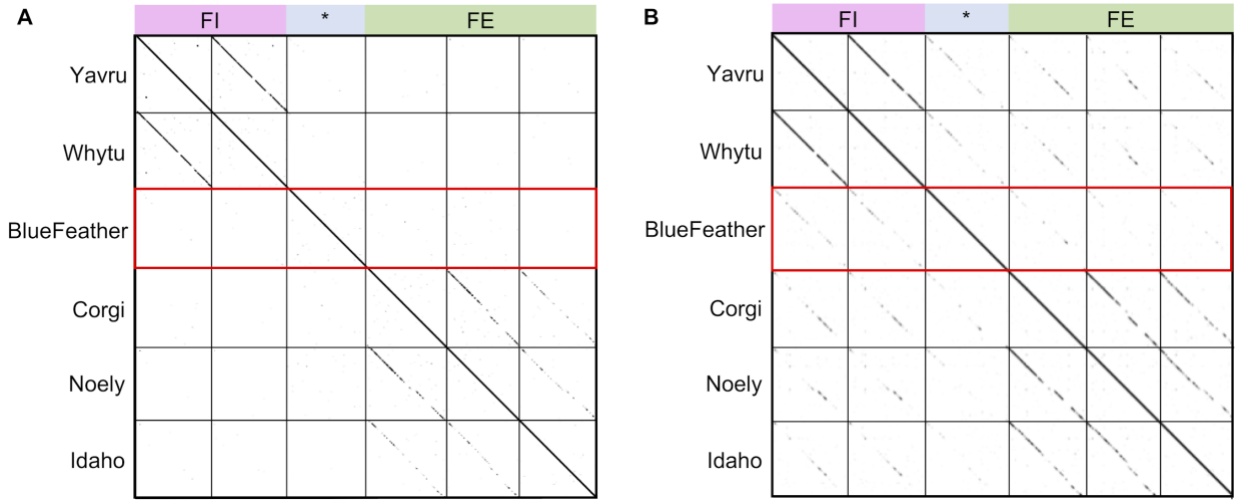


Fig 3. Dot plots suggest shared amino acids but not nucleotides. Whole genomes and proteomes for each phage were concatenated and dot plots were created using Gepard. Original cluster information is denoted along the top of each figure, with phage BlueFeather indicated by *. **(A)** A whole genome dot plot with word size of 15 indicates strong intracluster nucleotide similarities with both FE and former FI phages. No

intercluster nucleotide similarities were observed, indicating BlueFeather does not share significant nucleotide sequences with any of these phages. **(B)** A whole proteome dot plot with a word size of 5 indicated the same intracluster amino acid similarities seen in the genome dot plot, but there were also amino acid similarities observed between BlueFeather, Cluster FE, and former Cluster FI phages. BlueFeather appeared to have greater amino acid similarity with phages originally assigned to Cluster FI.

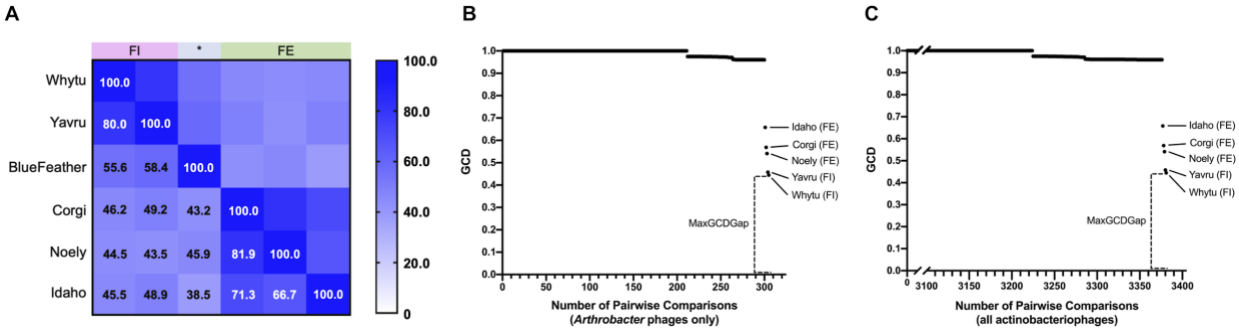


Fig 4. BlueFeather shares the most phams with phages originally assigned to Cluster FE and former Cluster FI (A) Gene Content Similarity (GCS) between BlueFeather, Cluster FE and the former Cluster FI was calculated with the PhagesDB GCS calculator using the number of shared phams. There was high intracluster GCS, and BlueFeather showed higher GCS values with former Cluster FI. **(B)** Gene Content Dissimilarity (GCD) output values of all pairwise comparisons of BlueFeather and all *Arthrobacter* phages (305), ordered by magnitude. Cluster FE and former Cluster FI were found to be least dissimilar to BlueFeather, with a MaxGCDGap of 44.45%, between BlueFeather and Whytu. **(C)** GCD output values of all pairwise comparisons of BlueFeather and all phages in PhagesDB (3381). MaxGCDGap remained at 44.45%. There are no non-*Arthrobacter* phages that are less dissimilar to BlueFeather than Whytu. BlueFeather shares up to 10% of genes with at least 63 non-*Arthrobacter* phages.

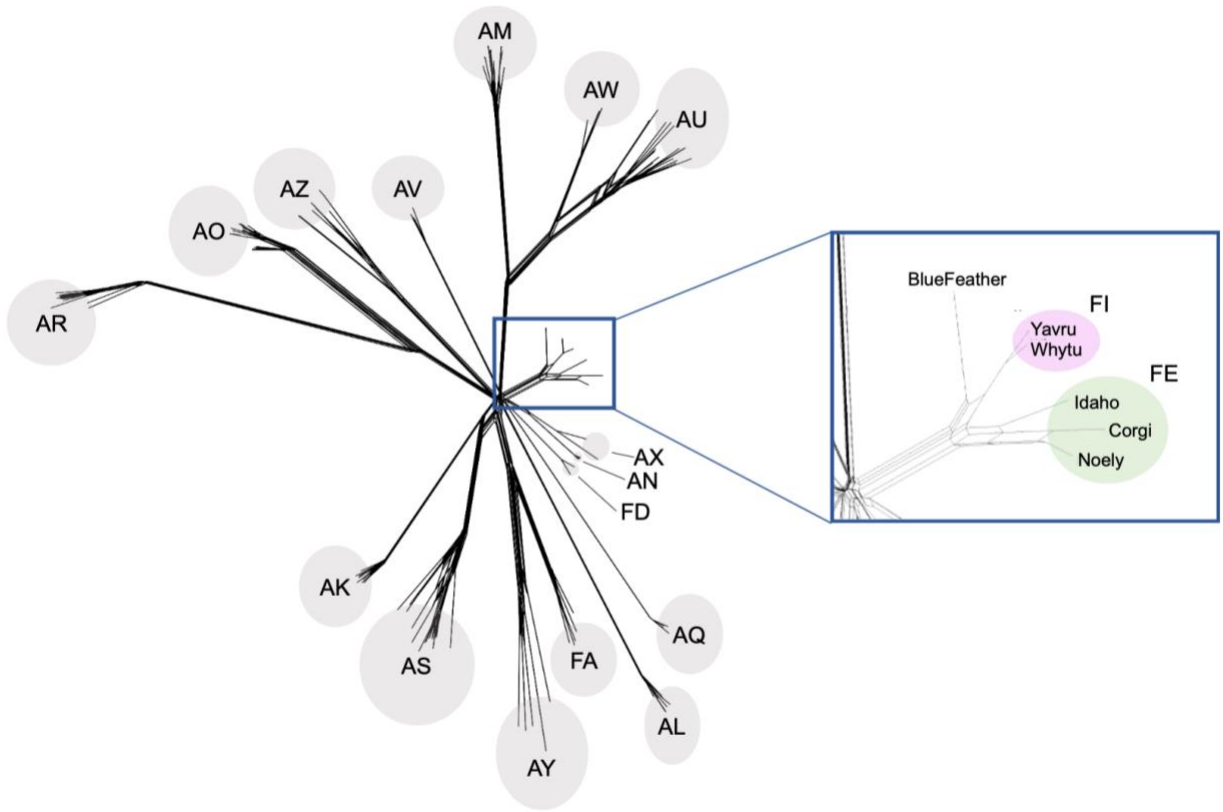


Fig 5. Expanded Cluster FE includes BlueFeather and former FI phages. A SplitsTree was generated in order to group *Arthrobacter* phages based on pham similarity. Ten representative phages from each cluster were selected to measure evolutionary relatedness. While there is great diversity of *Arthrobacter* phages, BlueFeather forms a relatively small branch with phages originally assigned to Cluster FE and the former Cluster FI. These phages, boxed in blue, comprise the expanded FE Cluster.

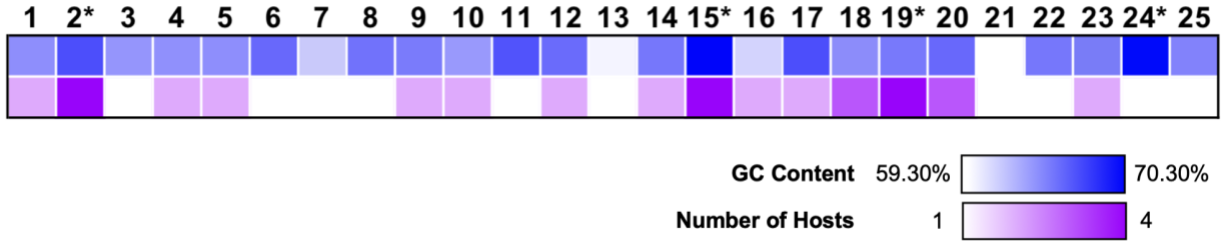


Fig 6. Evidence of horizontal gene transfer in the BlueFeather genome. The GC content for each gene in BlueFeather's genome ranged from 59.30%-70.30%, with an average of 64.30%. The number of unique isolation hosts that were represented in each pham ranged from 1-4. Genes with unexpectedly high values were considered to be the result of horizontal gene transfer. There were four genes with the most convincing evidence, indicated by *.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Besemer, John, and Mark Borodovsky. 2005. "GeneMark: Web Software for Gene Finding in Prokaryotes, Eukaryotes and Viruses." *Nucleic Acids Research* 33 (Web Server issue): W451-454. https://doi.org/10.1093/nar/gki487.

Bohlin, Jon, and John H.-O. Pettersson. 2019. "Evolution of Genomic Base Composition: From Single Cell Microbes to Multicellular Animals." *Computational and Structural Biotechnology Journal* 17: 362–70. https://doi.org/10.1016/j.csbj.2019.03.001.

Cresawn, Steven G., Matt Bogel, Nathan Day, Deborah Jacobs-Sera, Roger W. Hendrix, and Graham F. Hatfull. 2011. "Phamerator: A Bioinformatic Tool for Comparative Bacteriophage Genomics." *BMC Bioinformatics* 12: 395. https://doi.org/10.1186/1471-2105-12-395.

Delcher, Arthur L., Douglas Harmon, Simon Kasif, Owen White, and Steven L. Salzberg. 1999. "Improved Microbial Gene Identification with GLIMMER." *Nucleic Acids Research* 27 (23): 4636–41. https://doi.org/10.1093/nar/27.23.4636.

Grose, Julianne H., and Sherwood R. Casjens. 2014. "Understanding the Enormous Diversity of Bacteriophages: The Tailed Phages That Infect the Bacterial Family Enterobacteriaceae." *Virology* 468–470 (November): 421–43. https://doi.org/10.1016/j.virol.2014.08.024.

Hatfull, Graham F., Deborah Jacobs-Sera, Jeffrey G. Lawrence, Welkin H. Pope, Daniel A. Russell, Ching-Chung Ko, Rebecca J. Weber, et al. 2010. "Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size." *Journal of Molecular Biology* 397 (1): 119–43. https://doi.org/10.1016/j.jmb.2010.01.011.

Huson, D. H. 1998. "SplitsTree: Analyzing and Visualizing Evolutionary Data." *Bioinformatics (Oxford, England)* 14 (1): 68–73. https://doi.org/10.1093/bioinformatics/14.1.68.

Jacobs-Sera, Deborah, Lawrence A. Abad, Richard M. Alvey, Kirk R. Anders, Haley G. Aull, Suparna S. Bhalla, Lawrence S. Blumer, et al. 2020. "Genomic Diversity of Bacteriophages Infecting Microbacterium Spp." Edited by Axel Cloeckaert. *PLOS ONE* 15 (6): e0234636. https://doi.org/10.1371/journal.pone.0234636.

Keen, Eric C. 2015. "A Century of Phage Research: Bacteriophages and the Shaping of Modern Biology: Cause to Reflect." *BioEssays* 37 (1): 6–9. https://doi.org/10.1002/bies.201400152.

Klyczek, Karen K., J. Alfred Bonilla, Deborah Jacobs-Sera, Tamarah L. Adair, Patricia Afram, Katherine G. Allen, Megan L. Archambault, et al. 2017. "Tales of Diversity: Genomic and Morphological Characteristics of Forty-Six Arthrobacter Phages." Edited by Raymond Schuch. *PLOS ONE* 12 (7): e0180517. https://doi.org/10.1371/journal.pone.0180517.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305 (3): 567–80. https://doi.org/10.1006/jmbi.2000.4315.

Krumsiek, J., R. Arnold, and T. Rattei. 2007. "Gepard: A Rapid and Sensitive Tool for Creating Dotplots on Genome Scale." *Bioinformatics* 23 (8): 1026–28. https://doi.org/10.1093/bioinformatics/btm039.

Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, et al. 2014. "CDD: NCBI's Conserved Domain Database." *Nucleic Acids Research*, gku1221.

Mavrich, Travis N., and Graham F. Hatfull. 2017. "Bacteriophage Evolution Differs by Host, Lifestyle and Genome." *Nature Microbiology* 2 (9): 17112. https://doi.org/10.1038/nmicrobiol.2017.112.

Melo, Anna Cristhina Carmine de, Amanda da Mata Gomes, Fernando L. Melo, Daniel M. P. Ardisson-Araújo, Agueda Palmira Castagna de Vargas, Valessa Lunkes Ely, Elliot W. Kitajima, Bergmann M. Ribeiro, and José Luiz Caldas Wolff. 2019. "Characterization of a Bacteriophage with Broad Host Range against Strains of Pseudomonas Aeruginosa Isolated from Domestic Animals." *BMC Microbiology* 19 (1): 134. https://doi.org/10.1186/s12866-019-1481-z.

Miller-Ensminger, Taylor, Andrea Garretto, Jonathon Brenner, Krystal Thomas-White, Adriano Zambom, Alan J. Wolfe, and Catherine Putonti. 2018. "Bacteriophages of the Urinary Microbiome." *Journal of Bacteriology* 200 (7). https://doi.org/10.1128/JB.00738-17.

Ngandu, Nobubelo K, Konrad Scheffler, Penny Moore, Zenda Woodman, Darren Martin, and Cathal Seoighe. 2008. "Extensive Purifying Selection Acting on Synonymous Sites in HIV-1 Group M Sequences." *Virology Journal* 5 (1): 160. https://doi.org/10.1186/1743-422X-5-160.

Pedulla, Marisa L., Michael E. Ford, Jennifer M. Houtz, Tharun Karthikeyan, Curtis Wadsworth, John A. Lewis, Debbie Jacobs-Sera, et al. 2003. "Origins of Highly Mosaic Mycobacteriophage Genomes." *Cell* 113 (2): 171–82. https://doi.org/10.1016/s0092-8674(03)00233-2.

Pope, Welkin H, Charles A Bowman, Daniel A Russell, Deborah Jacobs-Sera, David J Asai, Steven G Cresawn, William R Jacobs Jr, et al. 2015. "Whole Genome Comparison of a Large Collection of Mycobacteriophages Reveals a Continuum of Phage Genetic Diversity." Edited by Roberto Kolter. *ELife* 4 (April): e06416. https://doi.org/10.7554/eLife.06416.

Pope, Welkin H., and Deborah Jacobs-Sera. 2018. "Annotation of Bacteriophage Genome Sequences Using DNA Master: An Overview." In *Bacteriophages*, edited by Martha R.J. Clokie, Andrew M. Kropinski, and Rob Lavigne, 1681:217–29. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7343-9_16.

Pope, Welkin H., Travis N. Mavrich, Rebecca A. Garlena, Carlos A. Guerrero-Bustamante, Deborah Jacobs-Sera, Matthew T. Montgomery, Daniel A. Russell, Marcie H. Warner, Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES), and Graham F. Hatfull. 2017. "Bacteriophages of Gordonia Spp. Display a Spectrum of Diversity and Genetic Relationships." *MBio* 8 (4). https://doi.org/10.1128/mBio.01069-17.

Russell, Daniel A. 2018. "Sequencing, Assembling, and Finishing Complete Bacteriophage Genomes." In *Bacteriophages: Methods and Protocols, Volume 3*, edited by Martha R.J. Clokie, Andrew M. Kropinski, and Rob Lavigne, 109–25. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7343-9_9.

Russell, Daniel A., and Graham F. Hatfull. 2017. "PhagesDB: The Actinobacteriophage Database." *Bioinformatics (Oxford, England)* 33 (5): 784–86. https://doi.org/10.1093/bioinformatics/btw711.

Sahoo, Dinabandhu, N. Jusna Devi, N. Ngashangva, P. Momota, Y. Rojeena, and S. Indira Devi. 2019. "Draft Genome Sequence of Arthrobacter Globiformis Mrc11, an Antimicrobial Agent Isolated from a Khangkhui Cave Deposit." Edited by Catherine Putonti. *Microbiology Resource Announcements* 8 (11): MRA.01620-18, e01620-18. https://doi.org/10.1128/MRA.01620-18.

Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri. 2012. “NIH Image to ImageJ: 25 Years of Image Analysis.” *Nature Methods* 9 (7): 671–75. https://doi.org/10.1038/nmeth.2089.

SEA-PHAGES/Starterator. (2016) 2020. Python. SEA-PHAGES Program. https://github.com/SEA-PHAGES/starterator.

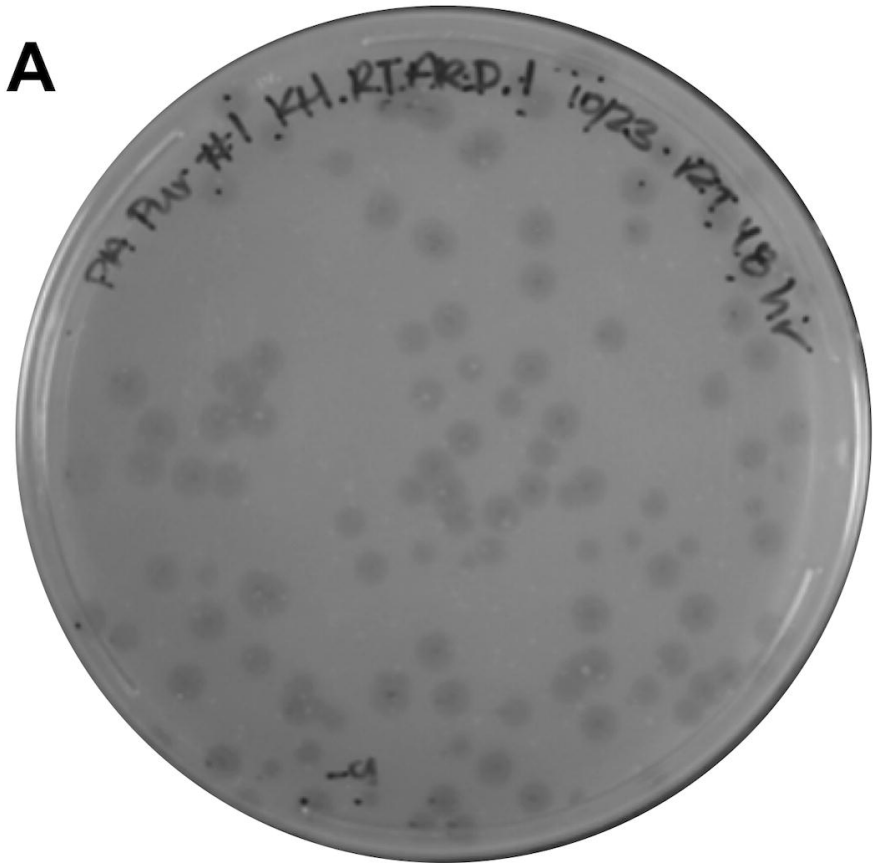
Shi, Yibo, Ning Li, Yaxian Yan, Hengan Wang, Yan Li, Chengping Lu, and Jianhe Sun. 2012. “Combined Antibacterial Activity of Phage Lytic Proteins Holin and Lysin from Streptococcus Suis Bacteriophage SMP.” *Current Microbiology* 65 (1): 28–34. https://doi.org/10.1007/s00284-012-0119-2.

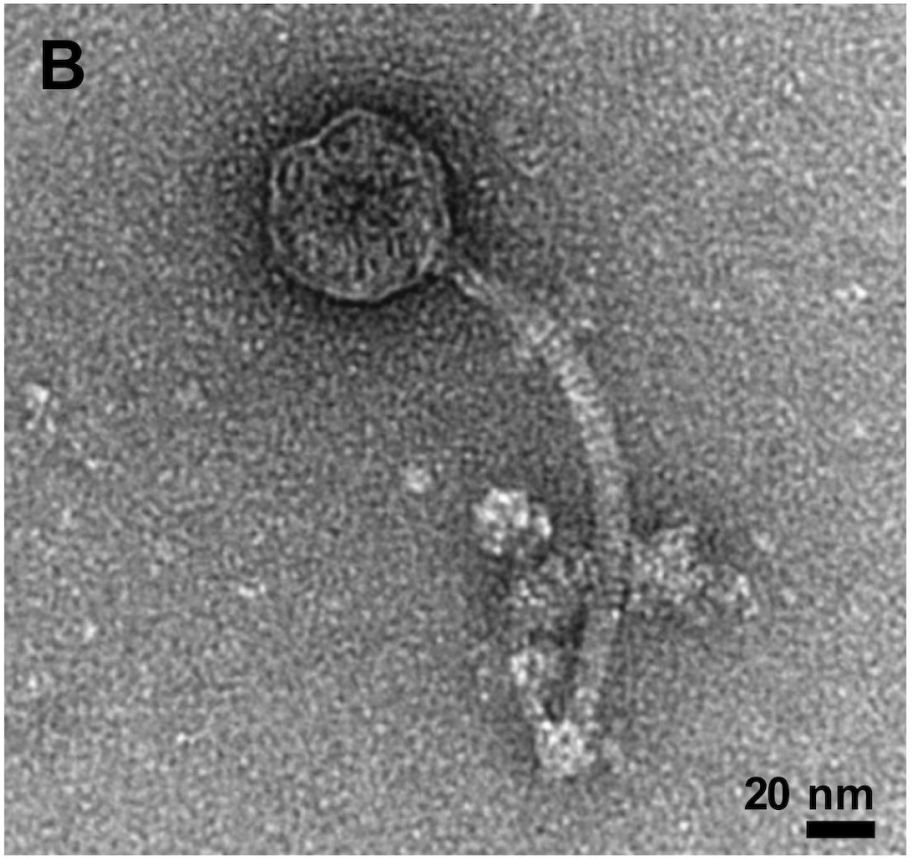
Soding, J., A. Biegert, and A. N. Lupas. 2005. “The HHpred Interactive Server for Protein Homology Detection and Structure Prediction.” *Nucleic Acids Research* 33 (Web Server): W244–48. https://doi.org/10.1093/nar/gki408.

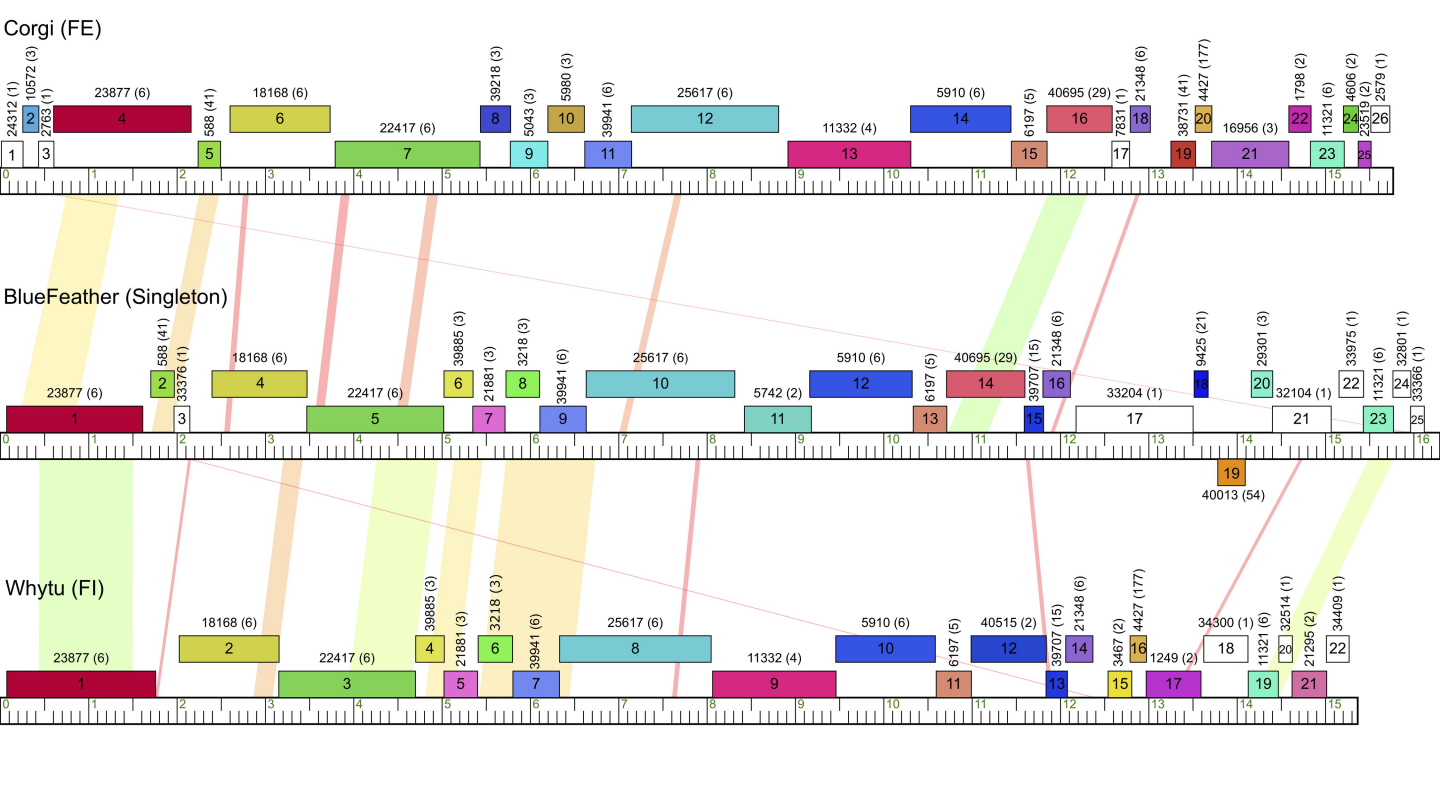
Tsirigos, Konstantinos D., Christoph Peters, Nanjiang Shu, Lukas Käll, and Arne Elofsson. 2015. “The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides.” *Nucleic Acids Research* 43 (W1): W401-407. https://doi.org/10.1093/nar/gkv485.

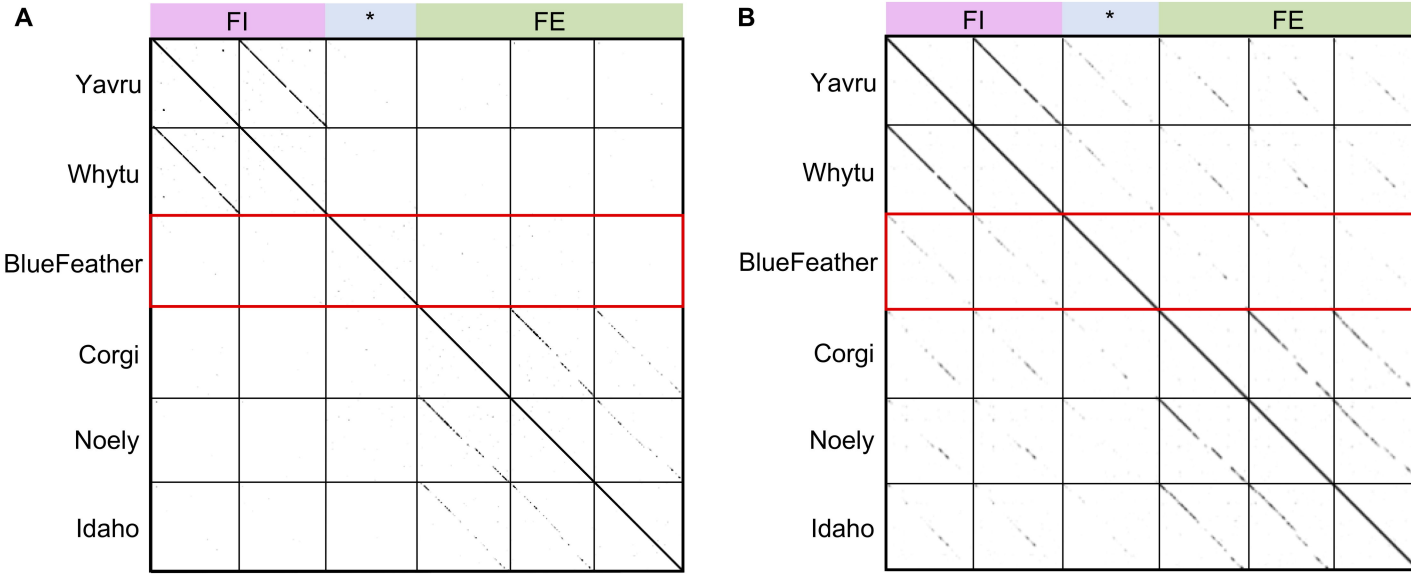
Wiles, Siouxsie. 2014. “Monday Micro – 200 Million Light Years of Viruses?!” Sciblogs. May 8, 2014. https://sciblogs.co.nz/infectious-thoughts/2014/08/05/monday-micro-200-million-light-years-of-viruses/.

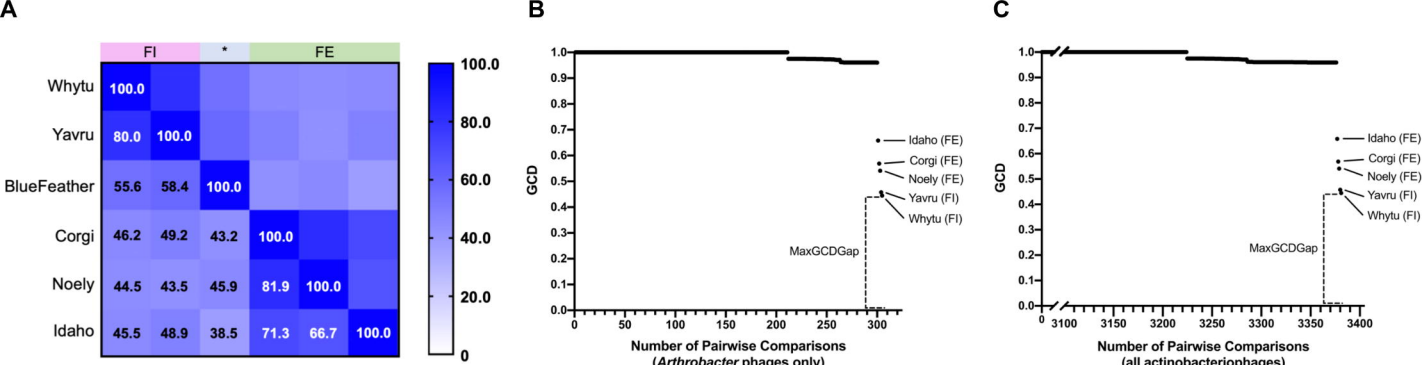
Yuan, Yihui, and Meiying Gao. 2017. “Jumbo Bacteriophages: An Overview.” *Frontiers in Microbiology* 8 (March). https://doi.org/10.3389/fmicb.2017.00403.

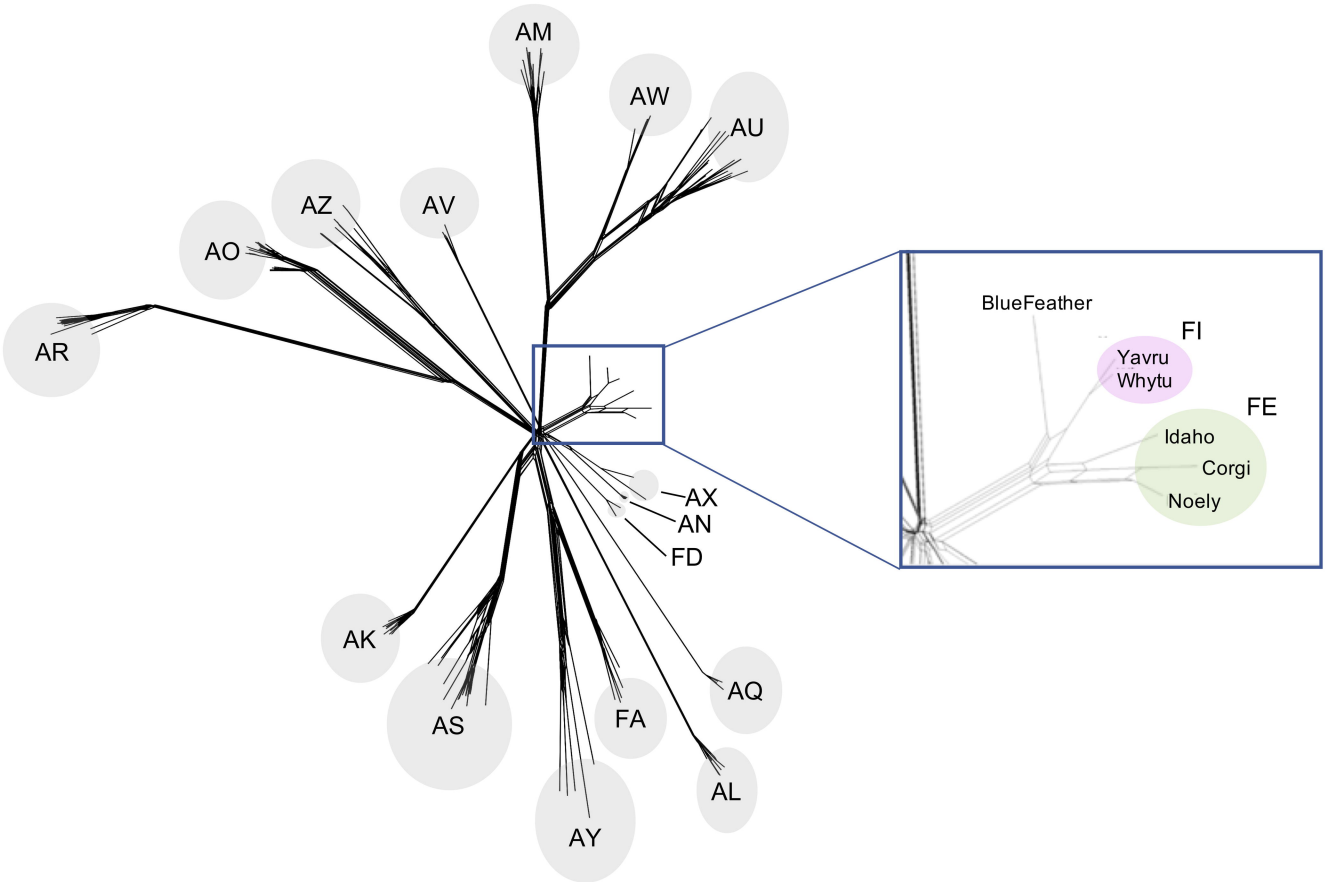

A

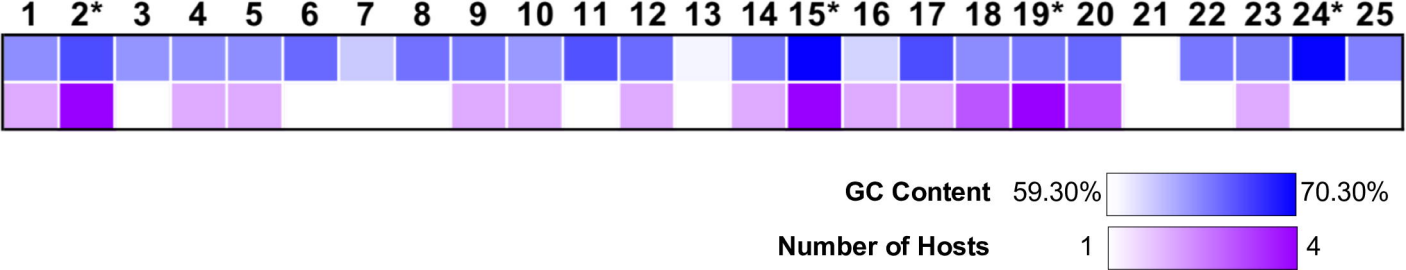

B
20 nm


Corgi (FE)
24312 (1)
10572 (3)
2763 (1)
23877 (6)
588 (41)
18168 (6)
22417 (6)
39218 (3)
5043 (3)
5980 (3)
39941 (6)
25617 (6)
11332 (4)
5910 (6)
6197 (5)
40695 (29)
7831 (1)
21348 (6)
38731 (41)
4427 (177)
16956 (3)
1798 (2)
11321 (6)
4606 (2)
23519 (2)
2579 (1)
BlueFeather (Singleton)
588 (41)
33376 (1)
18168 (6)
23877 (6)
22417 (6)
39885 (3)
21881 (3)
3218 (3)
39941 (6)
25617 (6)
5742 (2)
5910 (6)
6197 (5)
40695 (29)
39707 (15)
21348 (6)
9425 (21)
29301 (3)
33204 (1)
32104 (1)
33975 (1)
11321 (6)
32801 (1)
33366 (1)
40013 (54)
Whytu (FI)
18168 (6)
23877 (6)
22417 (6)
39885 (3)
21881 (3)
3218 (3)
39941 (6)
25617 (6)
11332 (4)
5910 (6)
6197 (5)
40515 (2)
39707 (15)
21348 (6)
3467 (2)
4427 (177)
1249 (2)
34300 (1)
11321 (6)
32514 (1)
21295 (2)
34409 (1)


A
FI
*
FE
Yavru
Whytu
BlueFeather
Corgi
Noely
Idaho
B
FI
*
FE
Yavru
Whytu
BlueFeather
Corgi
Noely
Idaho


A
FI
*
FE
Whytu
Yavru
BlueFeather
Corgi
Noely
Idaho
100.0
80.0 100.0
55.6 58.4 100.0
46.2 49.2 43.2 100.0
44.5 43.5 45.9 81.9 100.0
45.5 48.9 38.5 71.3 66.7 100.0
100.0
80.0
60.0
40.0
20.0
0
B
GCD
Idaho (FE)
Corgi (FE)
Noely (FE)
Yavru (FI)
Whytu (FI)
MaxGCDGap
Number of Pairwise Comparisons
(Arthrobacter phages only)
C
GCD
Idaho (FE)
Corgi (FE)
Noely (FE)
Yavru (FI)
Whytu (FI)
MaxGCDGap
Number of Pairwise Comparisons
(all actinobacteriophages)


AM
AW
AU
AZ
AV
AO
AR
AX
AN
FD
AK
AS
FA
AQ
AL
AY
BlueFeather
FI
Yavru
Whytu
FE
Idaho
Corgi
Noely


1
2*
3
4
5
6
7
8
9
10
11
12
13
14
15*
16
17
18
19*
20
21
22
23
24*
25
GC Content
59.30%
70.30%
Number of Hosts
1
4