bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Singleton phage BlueFeather reveals novel relationships in the *Arthrobacter* phage population

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

Singleton *Arthrobacter* phage BlueFeather

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Bacteriophages (phages) exhibit high genetic diversity, and the mosaic nature of the shared genetic pool makes quantifying phage relatedness an increasingly difficult task. Singleton phages lack the nucleotide identity and/or shared genes required for clustering with known phages. The goal of this study was to investigate singleton *Arthrobacter* phage BlueFeather's relationships with similar phages that had been assigned to clusters. Whole genome metrics showed low nucleotide identity, but high amino acid and gene content similarity between BlueFeather and *Arthrobacter* phage Clusters FE and FI. Gene content similarity revealed that BlueFeather shared genes with Clusters FE and FI in excess of a 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.

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 (1). Phages are found in all ecosystems in which bacteria exist and function as drivers of bacterial evolution (2). They exhibit horizontal gene transfer (HGT) with each other and with bacteria, resulting in the diverse and mosaic nature of phage genomes (3). Despite their incredible prevalence in the environment, phages remain largely understudied (4).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (5). 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 the cluster to warrant membership (6,7). A mass scale study on *Gordonia* phages also identified a spectrum of genetic diversity, as the clusters did not have clear boundaries (8). 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 (8), and later for *Microbacterium* phages (9), 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 found to exchange genes more slowly than *Gordonia* phages, and the 50% nucleotide clustering parameter was considered sufficient at the time (8). 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 (7), and this parameter has since been considered sufficient to categorize the limited number of *Arthrobacter* phages.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 singleton *Arthrobacter* phage BlueFeather was examined for nucleotide and amino acid identity with other known phages. BlueFeather lacked any significant nucleotide conservation for clustering, but had notable amino acid conservation and shared gene content with other clustered *Arthrobacter* phages, suggesting it may not be as isolated as its singleton status implied. The findings of this research on singleton phage BlueFeather expand the growing body of knowledge of phage genomics and serve as additional pieces of the puzzle of studying phage taxonomy.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (10).

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 (11).

Gene annotation

Genomes were annotated as described previously (12) using DNA Master (http://cobamide2.bio.pitt.edu/) and PECAAN (https://pecaan.kbrinsgd.org/) for auto-annotation. GLIMMER (13) and GeneMark (14) were used to predict protein-coding regions along with their start and stop sites. Manual annotation was performed using Phamerator (15), Starterator (16), and host-trained and self-trained GeneMark coding potential maps to support or refute auto-annotation predictions (14). Gene functions were determined using PhagesDB BLAST (https://phagesdb.org/blastp/), NCBI BLAST (17), HHpred (18) and CDD (19). The presence of transmembrane proteins was determined using TMHMM (20) and TOPCONS (21). The annotated complete genome was deposited to GenBank under the accession number MT024867.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gene content comparisons

Phage genomes used in this study are available from phagesdb.org (22). Gepard was used to perform sequence analysis to identify regions of homology between nucleotide sequences or amino acid sequences of different phages (23). 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 (24). 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 FI and Cluster FE phages (15). Gene Content Dissimilarity (GCD) and maximum GCD gap (MaxGCDGap) were calculated using scripts described previously (8). 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 the phages with genes in the same phams as BlueFeather's (22). 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/).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* (25).

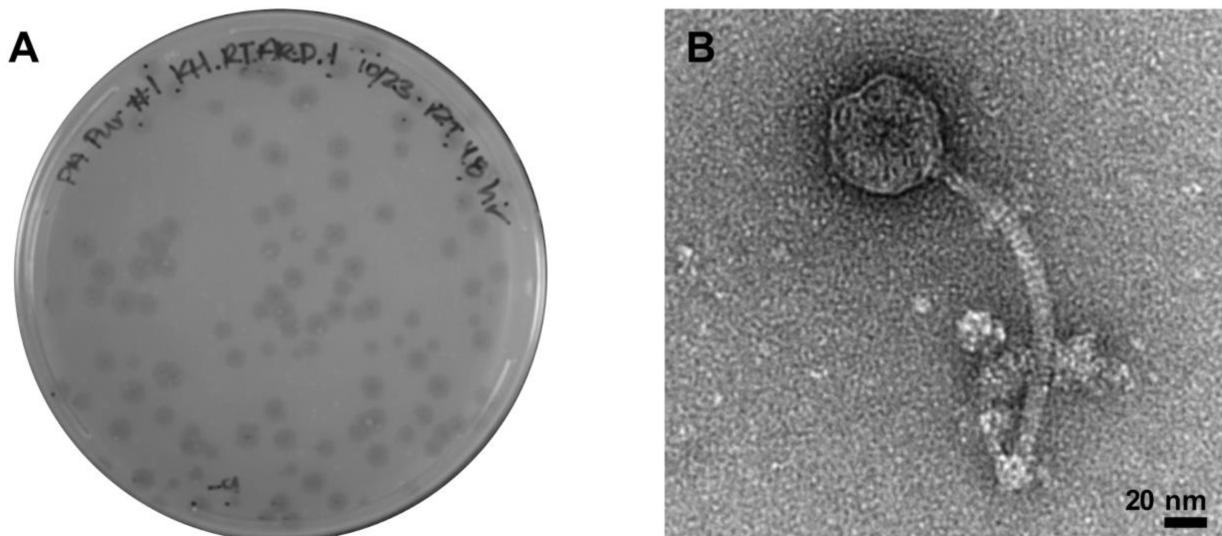


Fig 1. BlueFeather is a Siphovirus. (A) A picked plaque was suspended in phage buffer, then tested in a plaque assay for purification. Plaque morphology was consistent with bullseye plaques around 5 mm in diameter. **(B)** TEM image of BlueFeather at 67,000X magnification. The capsid was estimated to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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

70,265 bp (PrincessTrina) (7). 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 (7). 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 (26).

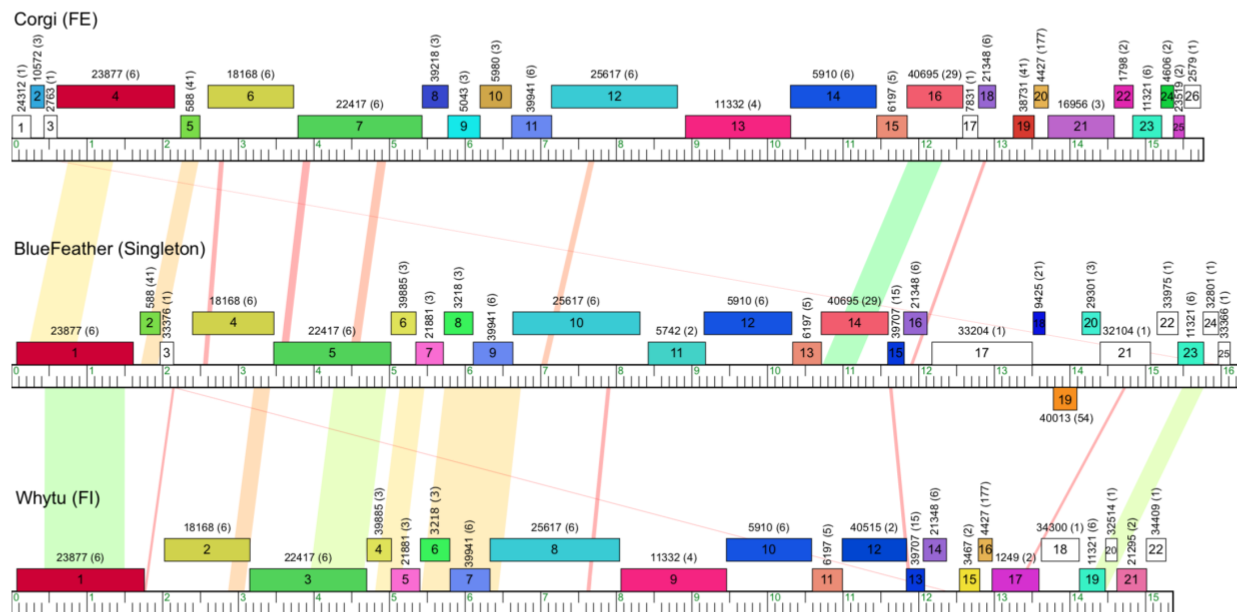


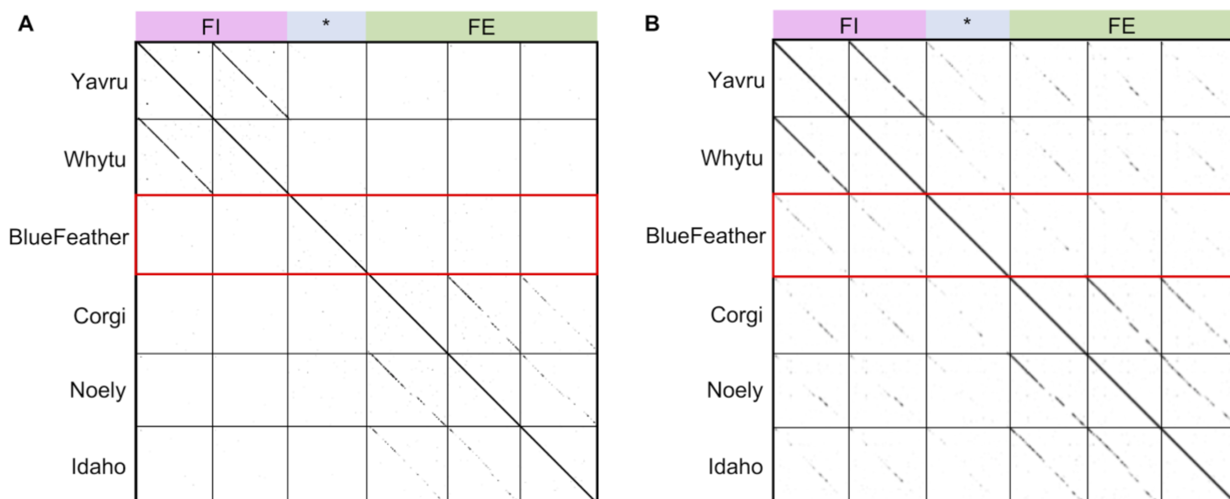
Fig 2. BlueFeather genome shares little nucleotide similarity but many phams with Clusters FE and FI. The BlueFeather genome is linear with a relatively small length of 16 kbp. Of the 25 identified ORFs, 18 were of known function, 6 were orphams

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Dot plot comparisons revealed synonymous substitutions in BlueFeather's genome

Phage BlueFeather was 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 phages in its most closely related *Arthrobacter* phage clusters, FE (Corgi, Idaho, Noely) and 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 known *Arthrobacter* clusters on PhagesDB, 17 clusters have between 2-4 phages. As expected, phages from the same clusters had alignments indicating large regions of nucleotide similarity (27), while comparison of BlueFeather's genome to those of Clusters FE and FI phages 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).



bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. Cluster information is denoted along the top of each figure, with singleton BlueFeather indicated by *. **a)** A whole genome dot plot with word size of 15 indicates strong intracluster nucleotide similarities with both FI and FE 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, Clusters FE and FI phages. BlueFeather appeared to have greater amino acid similarity with Cluster FI phages.

Gene similarity demonstrates a close relationship between BlueFeather and Clusters FE and 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 (28). GCS was calculated for BlueFeather and Clusters FE and FI phages. BlueFeather shared over 35% of genes with most Cluster FE phages, and over 55% of genes with 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 is a singleton, it was surprising to find GCS greater than the current threshold of 35% for clustering other phage populations (8,9). 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 (8). 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, FI, FH, and FF, 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, Clusters FE and FI phages 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.

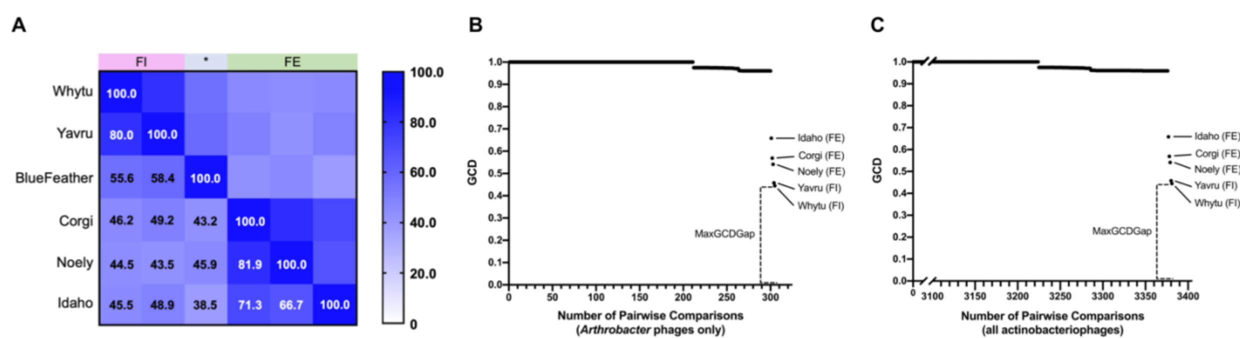


Fig 4. BlueFeather shares the most phams with Clusters FE and FI phages a) Gene Content Similarity (GCS) between BlueFeather, Cluster FE and Cluster FI was calculated with the PhagesDB GCS calculator using the number of shared phams. There was high intracluster GCS, and BlueFeather showed higher GCS values with Cluster FI. **b)** Gene Content Dissimilarity (GCD) output values of all pairwise comparisons of BlueFeather and all *Arthrobacter* phages (305), ordered by magnitude. Clusters FE and 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To compare the relationships between Clusters FE, FI, and BlueFeather with the *Arthrobacter* phage population as whole, 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 Clusters FE and FI than to any other *Arthrobacter* phage cluster (Fig 5). BlueFeather demonstrated a closer pham similarity to Cluster FI phages Whytu and Yavru than to Cluster FE phages Idaho, Noely and Corgi. Clusters FE, FI and BlueFeather formed a distinct branch from the rest of the phages sampled.

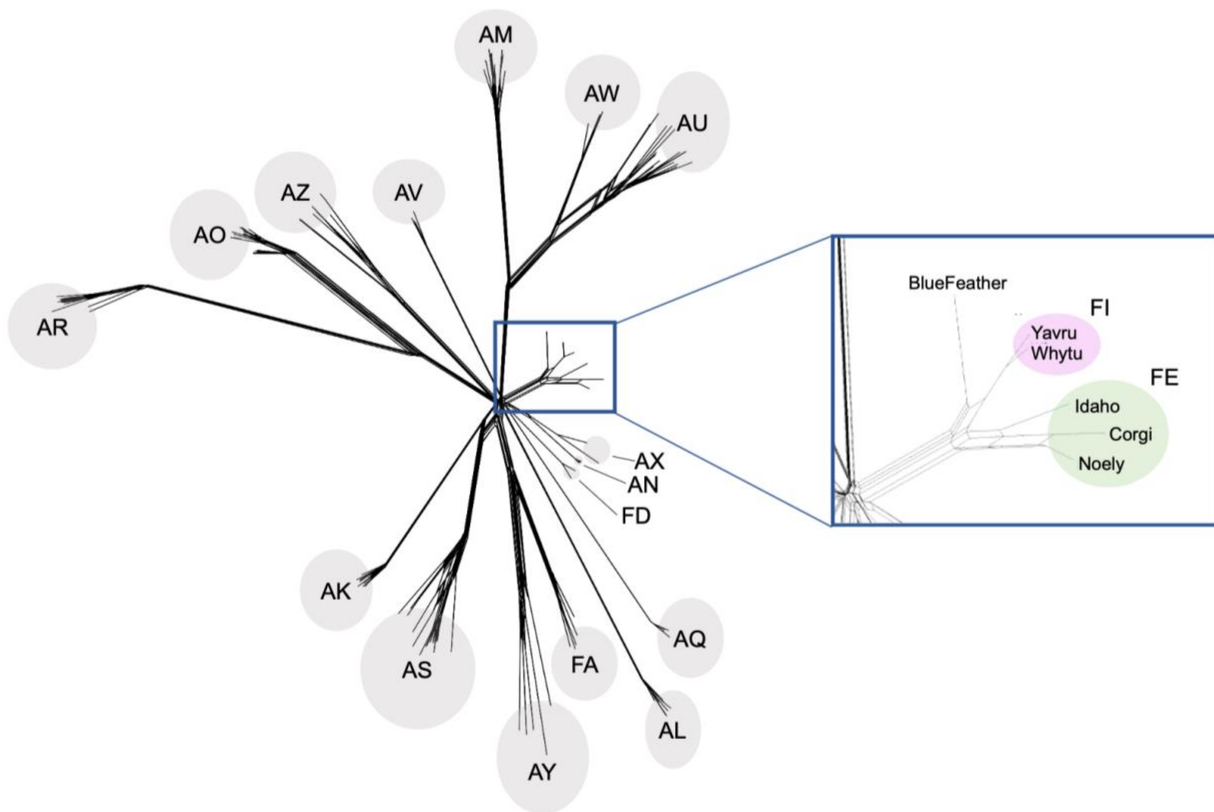


Fig 5. BlueFeather is most evolutionarily related to FI phages A SplitsTree was generated in order to group *Arthrobacter* phages based on pham similarity. Ten representative phages from each cluster were selected to measure evolutionary relatedness. While there is great diversity of *Arthrobacter* phages, BlueFeather forms a relatively small branch with Clusters FE and FI.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

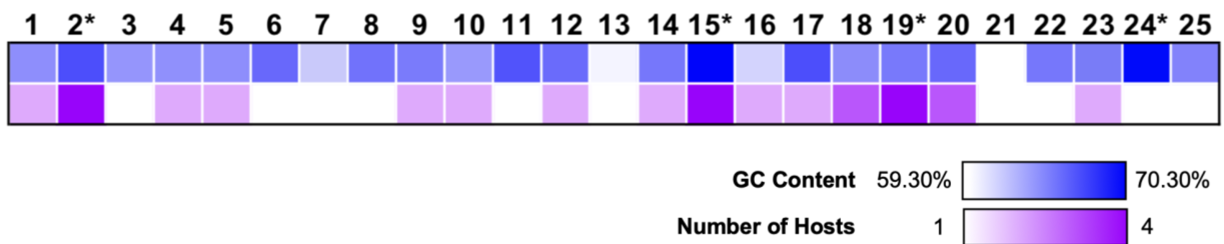


Fig 6. Evidence of horizontal gene transfer in the BlueFeather genome. The GC content for each gene in BlueFeather's genome ranged from 59.30%-70.30%, with an average of 64.30%. The number of unique isolation hosts that were represented in each pham ranged from 1-4. Genes with unexpectedly high values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

Typically, viral genes have about the same (29) or slightly lower GC content (30) 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% (31). 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

It is increasingly understood that phages infecting different hosts may share considerable gene content through processes such as HGT (8). 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* mc^{2}155, *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 novel singleton phage BlueFeather and other known phages, particularly those in 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 (7). BlueFeather, a singleton phage, exhibits over 35% GCS with nearly all phages from Cluster FE and over 55% GCS with those in Cluster FI, despite lacking the 50% nucleotide identity required for clustering. 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 (32). There is a high degree of synteny between these phages as well.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

While singleton status implies low relatedness to other phages (5), gene similarity between BlueFeather and phages from Clusters FE and FI in excess of 35% indicates conservation of gene functions and genome architecture despite extensive divergence of nucleotide identity. While *Arthrobacter* phages are currently clustered according to nucleotide identity (7), it seems that utilizing the shared gene content parameter could be beneficial to better capture the relatedness of BlueFeather and Clusters FE and FI phages, as is done for *Gordonia* and *Microbacterium* phages. BlueFeather has the smallest genome of all *Arthrobacter* singletons, and it is possible that clustering parameters may also need to be dependent on genome size. 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 Cluster FI phage Whytu. BlueFeather was found to be least dissimilar with Clusters FE and FI phages; this was supported by a network phylogeny of representative *Arthrobacter* phages that indicated great diversity between clusters, but revealed that singleton BlueFeather is most related to the Cluster FE and FI phages while remaining distinct. 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* phages to be very discrete (7) Even so, this low yet seemingly widespread display of shared genes, as well as BlueFeather's close relationships with Clusters FE and FI phages, provides new insight into the genetic

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

landscape of *Arthrobacter* phages. Clusters FE and FI are very small, containing only 5 of the 306 (as of May 8, 2020) sequenced and manually annotated *Arthrobacter* phages. On the other hand, there are 1,906 sequenced *Mycobacterium* phages (as of May 20, 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 other singleton phages that are replete with orphams (5,7), the BlueFeather genome is composed predominantly of genes with known functions that have been assigned to phams. BlueFeather has less than half as many genes as the other *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 (15).

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 (5). 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.

Instances in which there appears to be limited nucleotide conservation but high GCS, as observed with BlueFeather and Clusters FE and FI, support the notion that clustering methods should be continually reevaluated and optimized as more phages are sequenced (8). Further investigation into singleton phages is essential to entirely

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

understand the complex phage landscape. BlueFeather serves as yet another example of the highly intricate mosaic relationships which exist amongst phages, a feature of the genetic landscape which makes phage taxonomy an increasingly difficult task. The question is how to group *Arthrobacter* phages to best capture the continuum of diversity, as was previously done for *Gordonia* and *Microbacterium* phages (8,9). As more *Arthrobacter* phages are discovered, it is likely that we will discover many more phages like BlueFeather that are able to link discrete clusters by shared gene content. Clustering parameters are often designated according to the dataset they are meant to organize. As such, BlueFeather provides further evidence for the utility of clustering according to shared gene content, in order to best capture the mosaic nature of phage genomic relationships.

Acknowledgments

We thank Hong Zhou for electron microscopy assistance at the UCLA Electron Imaging Center for NanoMachines; Emily Pham and Oscar Lim for laboratory support; Rebecca A. Garlena and Daniel A. Russell at the Pittsburgh Bacteriophage Institute for genome sequence and assembly; and Travis Mavrich, Welkin Pope, and Graham Hatfull with the HHMI Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program for programmatic support.

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., and A.F. revised the paper; and J.M.P. and A.F. supervised the

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

References

1. Wiles S. Monday Micro – 200 million light years of viruses?! [Internet]. Sciblogs. 2014 [cited 2020 Aug 3]. Available from: https://sciblogs.co.nz/infectious-thoughts/2014/08/05/monday-micro-200-million-light-years-of-viruses/

2. Keen EC. A century of phage research: Bacteriophages and the shaping of modern biology: Cause to reflect. BioEssays. 2015 Jan;37(1):6–9.

3. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell. 2003 Apr 18;113(2):171–82.

4. Miller-Ensminger T, Garretto A, Brenner J, Thomas-White K, Zambom A, Wolfe AJ, et al. Bacteriophages of the Urinary Microbiome. J Bacteriol [Internet]. 2018 Mar 12 [cited 2020 Aug 10];200(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5847656/

5. Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Kolter R, editor. eLife. 2015 Apr 28;4:e06416.

6. Hatfull GF, Jacobs-Sera D, Lawrence JG, Pope WH, Russell DA, Ko C-C, et al. Comparative Genomic Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition, and Gene Size. J Mol Biol. 2010 Mar 19;397(1):119–43.

7. Klyczek KK, Bonilla JA, Jacobs-Sera D, Adair TL, Afram P, Allen KG, et al. Tales of diversity: Genomic and morphological characteristics of forty-six Arthrobacter phages. Schuch R, editor. PLOS ONE. 2017 Jul 17;12(7):e0180517.

8. Pope WH, Mavrich TN, Garlena RA, Guerrero-Bustamante CA, Jacobs-Sera D, Montgomery MT, et al. Bacteriophages of Gordonia spp. Display a Spectrum of Diversity and Genetic Relationships. mBio [Internet]. 2017 Sep 6 [cited 2020 Aug 10];8(4). Available from: https://mbio.asm.org/content/8/4/e01069-17

9. Jacobs-Sera D, Abad LA, Alvey RM, Anders KR, Aull HG, Bhalla SS, et al. Genomic diversity of bacteriophages infecting Microbacterium spp. Cloeckaert A, editor. PLOS ONE. 2020 Jun 18;15(6):e0234636.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012 Jul;9(7):671–5.

11. Russell DA. Sequencing, Assembling, and Finishing Complete Bacteriophage Genomes. In: Clokie MRJ, Kropinski AM, Lavigne R, editors. Bacteriophages: Methods and Protocols, Volume 3 [Internet]. New York, NY: Springer New York; 2018 [cited 2018 Nov 21]. p. 109–25. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-7343-9_9

12. Pope WH, Jacobs-Sera D. Annotation of Bacteriophage Genome Sequences Using DNA Master: An Overview. In: Clokie MRJ, Kropinski AM, Lavigne R, editors. Bacteriophages [Internet]. New York, NY: Springer New York; 2018 [cited 2020 Aug 10]. p. 217–29. (Methods in Molecular Biology; vol. 1681). Available from: http://link.springer.com/10.1007/978-1-4939-7343-9_16

13. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999 Dec 1;27(23):4636–41.

14. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W451-454.

15. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics. 2011;12:395.

16. SEA-PHAGES/starterator [Internet]. SEA-PHAGES Program; 2020 [cited 2020 Aug 10]. Available from: https://github.com/SEA-PHAGES/starterator

17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

18. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005 Jul 1;33(Web Server):W244–8.

19. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2014;gku1221.

20. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567–80.

21. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015 Jul 1;43(W1):W401-407.

22. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. Bioinforma Oxf Engl. 2017 01;33(5):784–6.

23. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007 Apr 15;23(8):1026–8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinforma Oxf Engl. 1998;14(1):68–73.

25. Yuan Y, Gao M. Jumbo Bacteriophages: An Overview. Front Microbiol [Internet]. 2017 Mar 14 [cited 2020 Aug 3];8. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00403/full

26. Shi Y, Li N, Yan Y, Wang H, Li Y, Lu C, et al. Combined Antibacterial Activity of Phage Lytic Proteins Holin and Lysin from Streptococcus suis Bacteriophage SMP. Curr Microbiol. 2012 Jul 1;65(1):28–34.

27. Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014 Nov;468–470:421–43.

28. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017 Sep;2(9):17112.

29. Bohlin J, Pettersson JH-O. Evolution of Genomic Base Composition: From Single Cell Microbes to Multicellular Animals. Comput Struct Biotechnol J. 2019;17:362–70.

30. de Melo ACC, da Mata Gomes A, Melo FL, Ardisson-Araújo DMP, de Vargas APC, Ely VL, et al. Characterization of a bacteriophage with broad host range against strains of Pseudomonas aeruginosa isolated from domestic animals. BMC Microbiol. 2019 Dec;19(1):134.

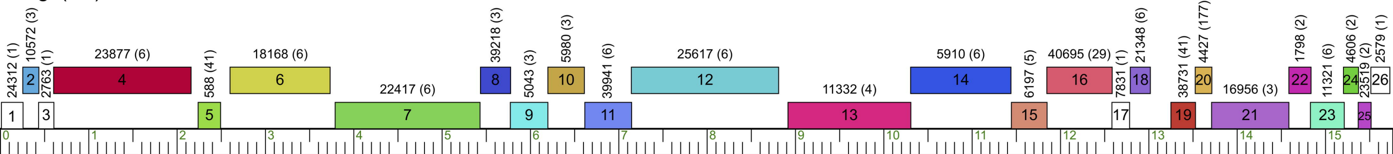
31. Sahoo D, Devi NJ, Ngashangva N, Momota P, Rojeena Y, Indira Devi S. Draft Genome Sequence of Arthrobacter globiformis mrc11, an Antimicrobial Agent Isolated from a Khangkhui Cave Deposit. Putonti C, editor. Microbiol Resour Announc. 2019 Mar 14;8(11):MRA.01620-18, e01620-18.

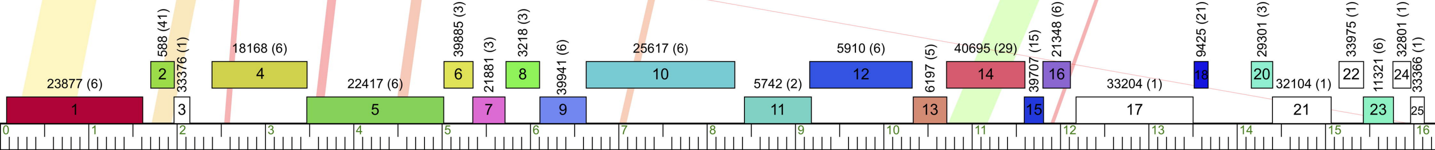
32. Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C. Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J. 2008;5(1):160.

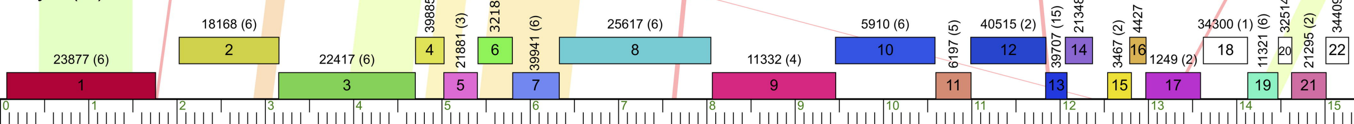
A

B

20 nm


Corgi (FE)


BlueFeather (Singleton)


Whytu (FI)

A

FI * FE

Yavru

Whytu

BlueFeather

Corgi

Noely

Idaho

B

FI * FE

Yavru

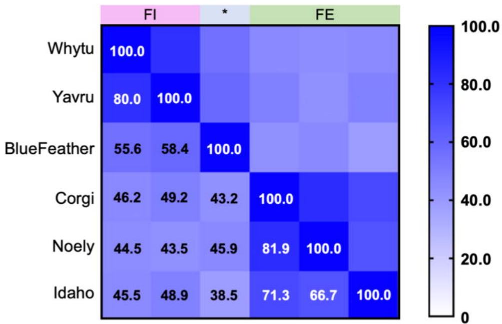
Whytu

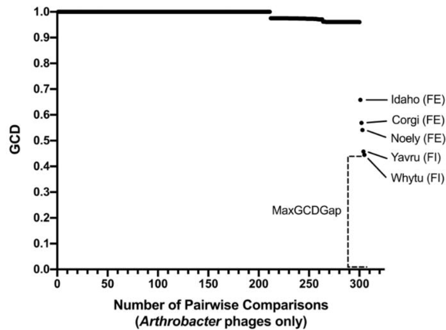
BlueFeather

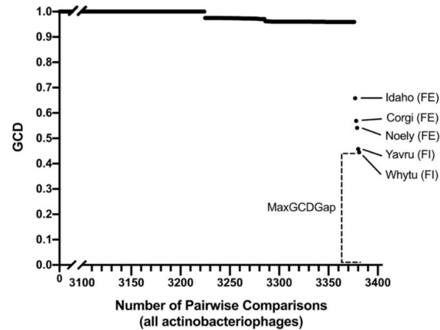
Corgi

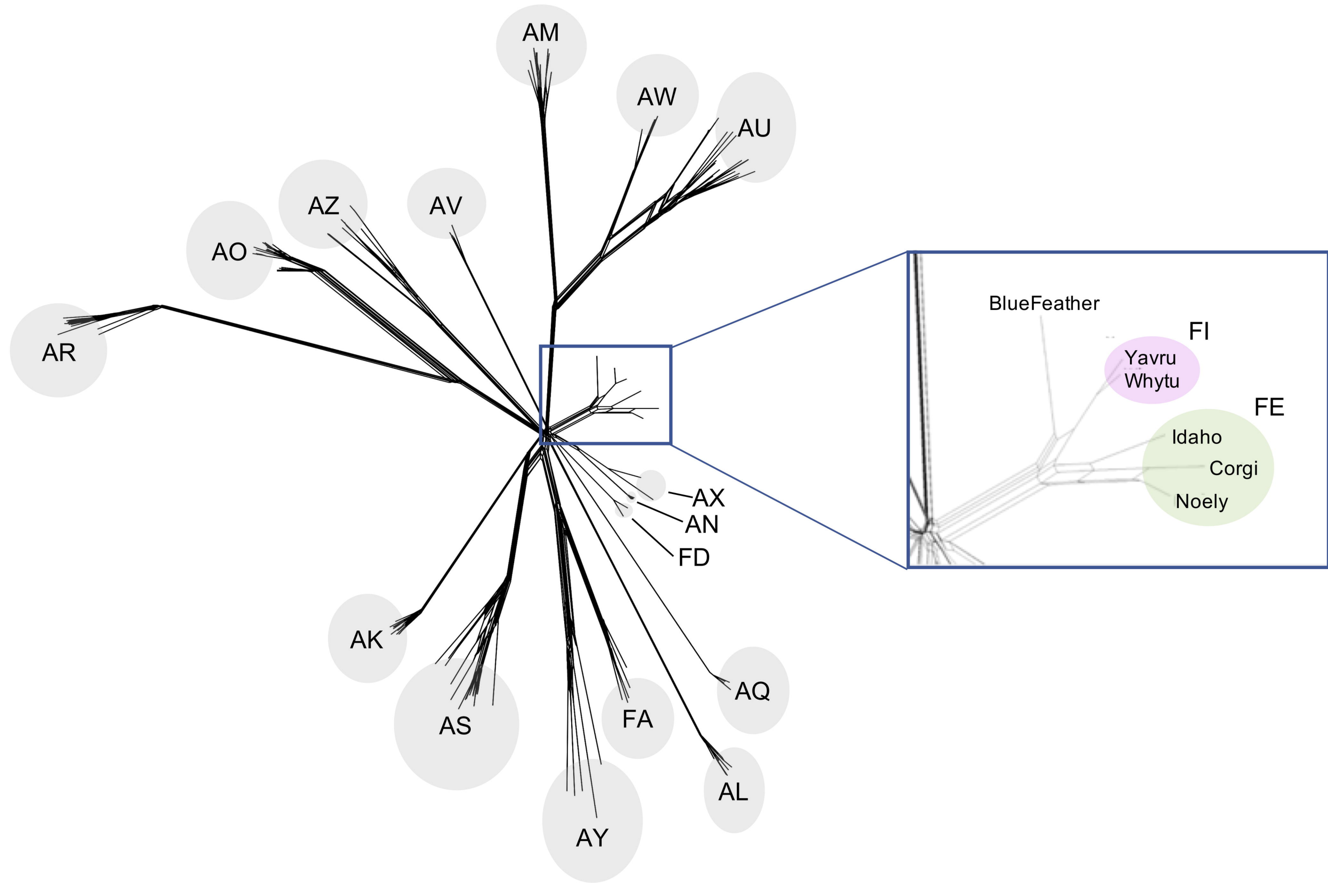
Noely

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
AY
FA
AL
AQ
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