bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

1 # BlueFeather, the singleton that wasn't: Shared gene
2 # content analysis supports expansion of *Arthrobacter*
3 # phage cluster FE
4
5

6 Stephanie Demo¶, Andrew Kapinos¶, Aaron Bernardino, Kristina Guardino, Blake
7 Hobbs, Kimberly Hoh, Edward Lee, Iphen Vuong, Krisanavane Reddi, Amanda C.
8 Freise, Jordan Moberg Parker*
9
10 ¶*Denotes equal contributions*
11
12 *Department of Microbiology, Immunology, and Molecular Genetics, University of*
13 *California, Los Angeles (UCLA), Los Angeles, CA, USA*
14

15 Keywords: bacteriophage, *Arthrobacter*, clustering, virology, microbiology, comparative
16 genomics
17

18 *Corresponding Author:
19 Jordan Moberg Parker, jmobergparker @ ucla.edu
20
21

22 **Brief Running Title**
23 *Arthrobacter* phage BlueFeather

24

25

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

26 # Abstract

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

44

45 # Introduction

46

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

51 2015). They exhibit horizontal gene transfer (HGT) with each other and with bacteria,
52 resulting in the diverse and mosaic nature of phage genomes (Pedulla et al. 2003).
53 Despite their incredible prevalence in the environment, phages remain largely
54 understudied (Miller-Ensminger et al. 2018).

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

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

92

93 # Materials and methods

94

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

117 ## Gene annotation

118 Genomes were annotated as described previously (Pope and Jacobs-Sera 2018)
119 using DNA Master (http://cobamide2.bio.pitt.edu/) and PECAAN
120 (https://pecaan.kbrinsgd.org/) for auto-annotation. GLIMMER (Delcher et al. 1999) and
121 GeneMark (Besemer and Borodovsky 2005) were used to predict protein-coding regions
122 along with their start and stop sites. Manual annotation was performed using
123 Phamerator (Cresawn et al. 2011), Starterator (*SEA-PHAGES/Starterator* [2016] 2020), and
124 host-trained and self-trained GeneMark coding potential maps to support or refute auto-
125 annotation predictions (Besemer and Borodovsky 2005). Gene functions were determined
126 using PhagesDB BLAST (https://phagesdb.org/blastp/), NCBI BLAST (Altschul et al.
127 1990), HHpred (Soding, Biegert, and Lupas 2005) and CDD (Marchler-Bauer et al. 2014). The
128 presence of transmembrane proteins was determined using TMHMM (Krogh et al. 2001)
129 and TOPCONS (Tsirigos et al. 2015). The annotated complete genome was deposited to
130 GenBank under the accession number MT024867.

131 ## Gene content comparisons

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

140 representative phages from each *Arthrobacter* cluster were selected from the
141 Actino_Draft database (version 366) for comparison.
142 The gene content calculator on PhagesDB (https://phagesdb.org/genecontent/)
143 was used to calculate Gene Content Similarity (GCS), the percentage of shared genes
144 in phams (groups of genes with related sequences), between BlueFeather, Cluster FE,
145 and former Cluster FI phages (Cresawn et al. 2011). Gene Content Dissimilarity (GCD)
146 and maximum GCD gap (MaxGCDGap) were calculated using scripts described
147 previously (Pope et al. 2017). Heatmaps and scatter plots were created using Prism 8.0.0
148 (GraphPad Software, San Diego, California, USA) and were used for quantitative
149 analysis and visualization of GCS and GCD values.
150 PhagesDB Pham View was used to gather information about phages with genes
151 in the same phams as BlueFeather's (Russell and Hatfull 2017). PECAAN was used to
152 obtain the nucleotide sequences for each BlueFeather gene
153 (https://discover.kbrinsgd.org). The BiologicsCorp online GC content calculator was
154 used for each gene in the genome (https://www.biologicscorp.com/tools/GCContent/).

155 Results

156 BlueFeather is a siphovirus with a short genome

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

163 BlueFeather's genome had a length of 16,302 bp, 64.30% GC content, and
164 genome ends with 15 base 3' sticky overhangs (CCACGGTTCCCGTCC). Phages that
165 infect *Arthrobacter* hosts have genome lengths that range from 15,319 bp (Toulouse) to
166 70,265 bp (PrincessTrina) (Klyczek et al. 2017). The average *Arthobacter* phage genome
167 length (as of May 2020) was 46,968 bp with a standard deviation of 20,619 bp and a
168 median length of 53,859 bp, suggesting that most *Arthrobacter* phages have genomes
169 notably larger than that of BlueFeather. BlueFeather's genome contained 25 manually
170 annotated genes; 18 were of known function, 6 were orphams–meaning they have not
171 been identified in any other known phage–and 1 was a reverse gene (Fig 2). The left
172 arm of the genome had highly conserved genes amongst siphoviral *Arthrobacter*
173 phages, such as those encoding terminase, portal protein, head-to-tail adapter, and tail
174 proteins (Klyczek et al. 2017). Tail tube and sheath genes were absent, confirming the
175 classification of BlueFeather as a siphovirus. Genes characteristic of the lytic life cycle,
176 such as lysin A and holin, were identified; however, there were no genes that would
177 indicate BlueFeather's ability to undergo a lysogenic life cycle, suggesting that
178 BlueFeather is not a temperate phage (Shi et al. 2012).

179 Dot plot comparisons revealed synonymous substitutions in 180 BlueFeather's genome

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

196 Gene similarity demonstrates a close relationship between 197 BlueFeather and phages of Cluster FE and former Cluster FI

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

235 BlueFeather genome exhibits evidence of horizontal gene 236 transfer

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

243 Typically, viral genes have about the same (Bohlin and Pettersson 2019) or slightly
244 lower GC content (de Melo et al. 2019) compared to their bacterial hosts, suggesting that
245 genes with higher GC content may have been horizontally transferred. BlueFeather had
246 an overall average GC content of 64.30% and *Arthrobacter globiformis* mrc11 was
247 found to have an overall GC content of 65.9% (Sahoo et al. 2019). BlueFeather gene-
248 specific average GC contents ranged from 59.30% to 70.30%, and genes with
249 maximum average GC contents were considered for HGT. This included genes 15 and
250 24 with GC contents of 70.3% and 70.1%, respectively.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

255 member genes found in phages that infect *Gordonia malaquae* BEN700 and
256 *Arthrobacter sp.* ATCC 21022. Gene 15 belongs to a pham with member genes found in
257 phages that infect *A. globiformis* B-2979, *A. sp*. ATCC 21022, *Mycobacterium*
258 *smegmatis* mc²155, *G. malaquae* BEN700, and *Gordonia rubripertincta* NRRL B-16540.
259 Gene 19 was the only reverse gene in the BlueFeather genome, and this gene was only
260 found in BlueFeather and in phages infecting *Microbacterium foliorum* NRRL B-24224
261 SEA and *Microbacterium paraoxydans* NWU1.
262

263 # Discussion

264 Our research was focused on the genomic and evolutionary relationships
265 between the novel *Arthrobacter* phage BlueFeather and other known phages,
266 particularly those originally assigned to Clusters FE and FI. Previous studies have
267 shown that new clusters can be formed when novel phages are found to be similar to
268 former singletons, as demonstrated by the formation of Cluster AS from *Arthrobacter*
269 singleton Galaxy (Klyczek et al. 2017). BlueFeather, originally designated as a putative
270 singleton phage, exhibits over 35% GCS with nearly all phages originally in Cluster FE
271 and over 55% GCS with those formerly assigned to Cluster FI. The conservation of
272 amino acids, rather than nucleotides, suggests a history of purifying selection via many
273 synonymous mutations in which deleterious mutations were filtered out (Ngandu et al.
274 2008). Moreover, this presents some of the most clear evidence to date of highly
275 conserved amino acid sequences despite the absence of significant nucleotide
276 conservation amongst related phages.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

291 Gene content dissimilarity demonstrated that BlueFeather has a MaxGCDGap of
292 44.45% with phage Whytu, which was originally assigned to Cluster FI. BlueFeather
293 was found to be least dissimilar with Cluster FE and former Cluster FI phages; this was
294 supported by a network phylogeny of representative *Arthrobacter* phages that indicated
295 great diversity between clusters, but revealed that phage BlueFeather forms a distinct
296 branch with Cluster FE and former Cluster FI phages. Additionally, many phages were
297 found to share between 0-10% GCS with BlueFeather. While this is too low to warrant a
298 significant phylogenetic relationship, it reinforced the observed continuum of diversity in
299 phage populations. Previous research found *Arthrobacter* phage clusters to be very

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

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

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

343

344 Acknowledgments

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

347 Welkin Pope, Debbie Jacobs-Sera, and Graham Hatfull with the HHMI Science
348 Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science
349 (SEA-PHAGES) program for programmatic support. The authors acknowledge the use
350 of instruments at the Electron Imaging Center for NanoMachines supported by NIH
351 (1S10RR23057 to ZHZ) and CNSI at UCLA.

352

353 ## Author contributions

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

358

359 ## Author disclosure statement

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

362

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

Figures

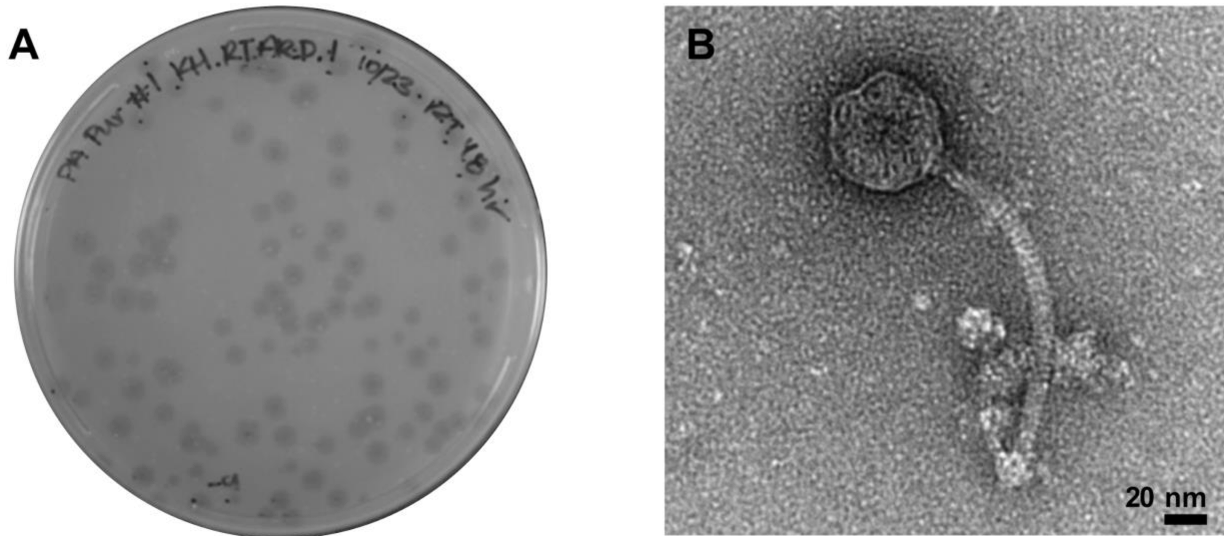


Fig 1. BlueFeather is a Siphovirus. (A) A picked plaque was suspended in phage buffer, then tested in a plaque assay for purification. Plaque morphology was consistent with bullseye plaques around 5 mm in diameter. **(B)** TEM image of BlueFeather at 67,000X magnification. The capsid was estimated to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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

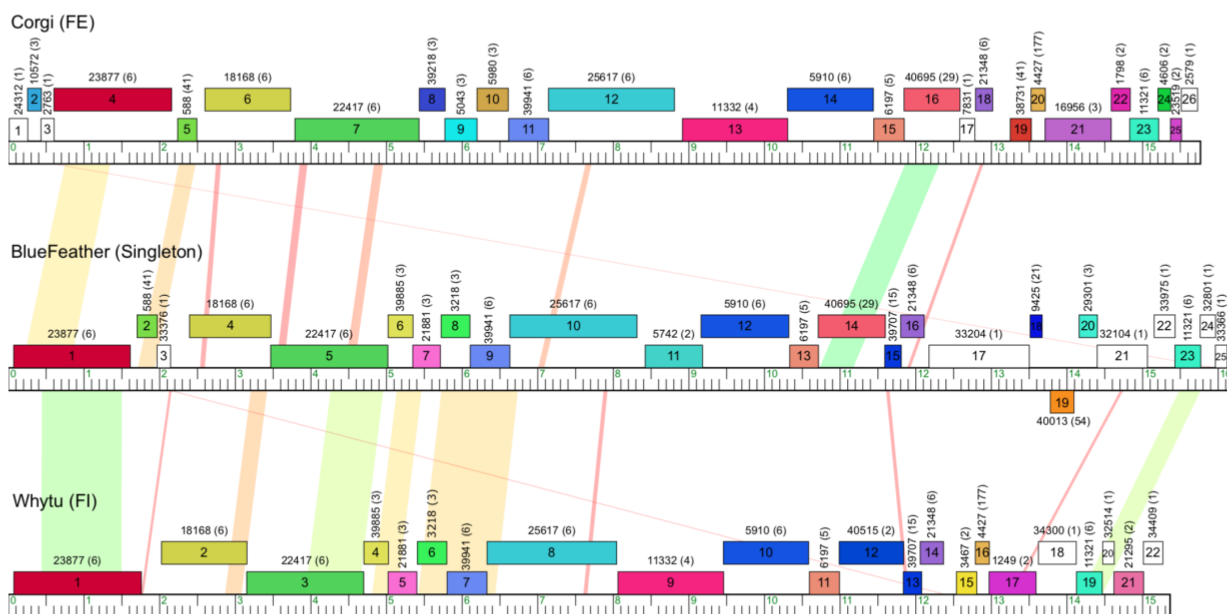


Fig 2. BlueFeather genome shares little nucleotide similarity but many phams with Cluster FE and former Cluster FI. The BlueFeather genome is linear with a relatively small length of 16 kbp. Of the 25 identified ORFs, 18 were of known function, 6 were orphams and 1 was a reverse gene. BlueFeather had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

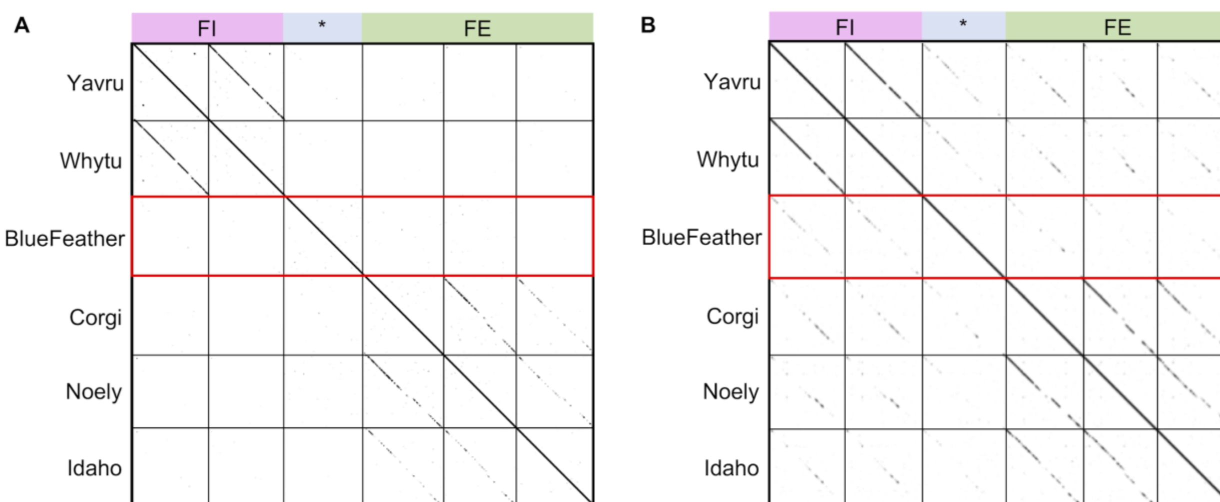
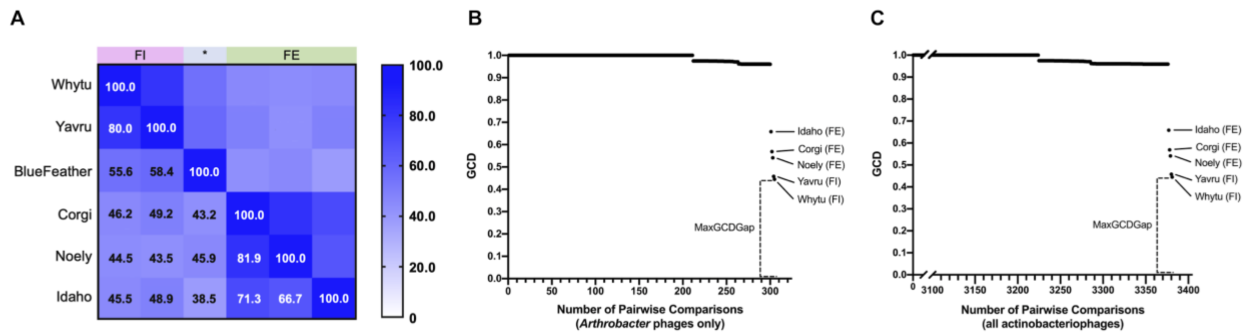


Fig 3. Dot plots suggest shared amino acids but not nucleotides. Whole genomes and proteomes for each phage were concatenated and dot plots were created using Gepard. Original cluster information is denoted along the top of each figure, with phage BlueFeather indicated by *. **(A)** A whole genome dot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

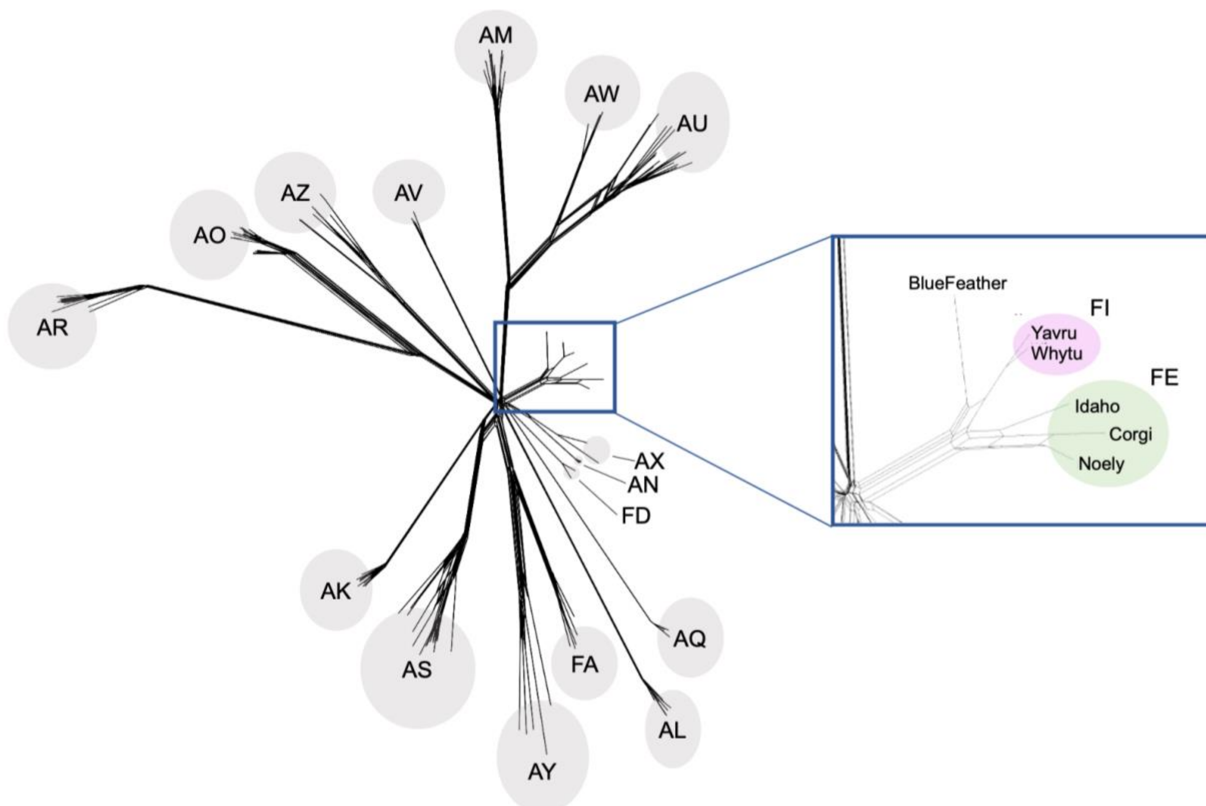
bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

387 intercluster nucleotide similarities were observed, indicating BlueFeather does not share
388 significant nucleotide sequences with any of these phages. **(B)** A whole proteome dot
389 plot with a word size of 5 indicated the same intracluster amino acid similarities seen in
390 the genome dot plot, but there were also amino acid similarities observed between
391 BlueFeather, Cluster FE, and former Cluster FI phages. BlueFeather appeared to have
392 greater amino acid similarity with phages originally assigned to Cluster FI.
393

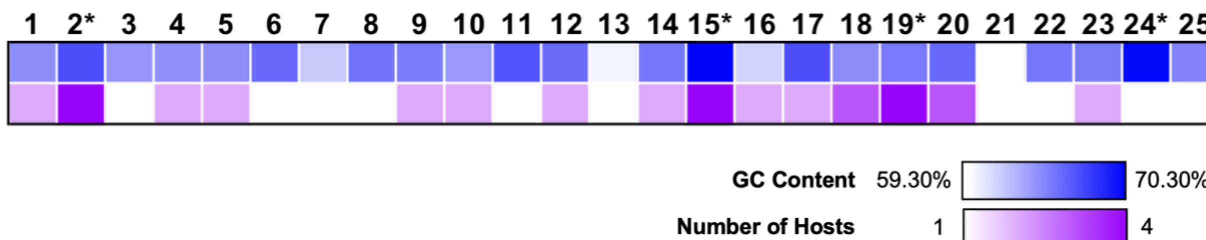


394
395
396
397 **Fig 4. BlueFeather shares the most phams with phages originally assigned to**
398 **Cluster FE and former Cluster FI (A)** Gene Content Similarity (GCS) between
399 BlueFeather, Cluster FE and the former Cluster FI was calculated with the PhagesDB
400 GCS calculator using the number of shared phams. There was high intracluster GCS,
401 and BlueFeather showed higher GCS values with former Cluster FI. **(B)** Gene Content
402 Dissimilarity (GCD) output values of all pairwise comparisons of BlueFeather and all
403 *Arthrobacter* phages (305), ordered by magnitude. Cluster FE and former Cluster FI
404 were found to be least dissimilar to BlueFeather, with a MaxGCDGap of 44.45%,
405 between BlueFeather and Whytu. **(C)** GCD output values of all pairwise comparisons of
406 BlueFeather and all phages in PhagesDB (3381). MaxGCDGap remained at 44.45%.
407 There are no non-*Arthrobacter* phages that are less dissimilar to BlueFeather than
408 Whytu. BlueFeather shares up to 10% of genes with at least 63 non-*Arthrobacter*
409 phages.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.



410
411
412 **Fig 5. Expanded Cluster FE includes BlueFeather and former FI phages.** A
413 SplitsTree was generated in order to group *Arthrobacter* phages based on pham
414 similarity. Ten representative phages from each cluster were selected to measure
415 evolutionary relatedness. While there is great diversity of *Arthrobacter* phages,
416 BlueFeather forms a relatively small branch with phages originally assigned to Cluster
417 FE and the former Cluster FI. These phages, boxed in blue, comprise the expanded FE
418 Cluster.
419
420



421
422 **Fig 6. Evidence of horizontal gene transfer in the BlueFeather genome.** The GC
423 content for each gene in BlueFeather's genome ranged from 59.30%-70.30%, with an
424 average of 64.30%. The number of unique isolation hosts that were represented in each
425 pham ranged from 1-4. Genes with unexpectedly high values were considered to be the
426 result of horizontal gene transfer. There were four genes with the most convincing
427 evidence, indicated by *.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

428
429
430 # References
431

432 Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990.
433 "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10.
434 https://doi.org/10.1016/S0022-2836(05)80360-2.
435 Besemer, John, and Mark Borodovsky. 2005. "GeneMark: Web Software for Gene Finding in
436 Prokaryotes, Eukaryotes and Viruses." *Nucleic Acids Research* 33 (Web Server issue):
437 W451-454. https://doi.org/10.1093/nar/gki487.
438 Bohlin, Jon, and John H.-O. Pettersson. 2019. "Evolution of Genomic Base Composition: From
439 Single Cell Microbes to Multicellular Animals." *Computational and Structural*
440 *Biotechnology Journal* 17: 362–70. https://doi.org/10.1016/j.csbj.2019.03.001.
441 Cresawn, Steven G., Matt Bogel, Nathan Day, Deborah Jacobs-Sera, Roger W. Hendrix, and
442 Graham F. Hatfull. 2011. "Phamerator: A Bioinformatic Tool for Comparative
443 Bacteriophage Genomics." *BMC Bioinformatics* 12: 395. https://doi.org/10.1186/1471-
444 2105-12-395.
445 Delcher, Arthur L., Douglas Harmon, Simon Kasif, Owen White, and Steven L. Salzberg. 1999.
446 "Improved Microbial Gene Identification with GLIMMER." *Nucleic Acids Research* 27
447 (23): 4636–41. https://doi.org/10.1093/nar/27.23.4636.
448 Grose, Julianne H., and Sherwood R. Casjens. 2014. "Understanding the Enormous Diversity of
449 Bacteriophages: The Tailed Phages That Infect the Bacterial Family
450 Enterobacteriaceae." *Virology* 468–470 (November): 421–43.
451 https://doi.org/10.1016/j.virol.2014.08.024.
452 Hatfull, Graham F., Deborah Jacobs-Sera, Jeffrey G. Lawrence, Welkin H. Pope, Daniel A.
453 Russell, Ching-Chung Ko, Rebecca J. Weber, et al. 2010. "Comparative Genomic
454 Analysis of 60 Mycobacteriophage Genomes: Genome Clustering, Gene Acquisition,
455 and Gene Size." *Journal of Molecular Biology* 397 (1): 119–43.
456 https://doi.org/10.1016/j.jmb.2010.01.011.
457 Huson, D. H. 1998. "SplitsTree: Analyzing and Visualizing Evolutionary Data." *Bioinformatics*
458 *(Oxford, England)* 14 (1): 68–73. https://doi.org/10.1093/bioinformatics/14.1.68.
459 Jacobs-Sera, Deborah, Lawrence A. Abad, Richard M. Alvey, Kirk R. Anders, Haley G. Aull,
460 Suparna S. Bhalla, Lawrence S. Blumer, et al. 2020. "Genomic Diversity of
461 Bacteriophages Infecting Microbacterium Spp." Edited by Axel Cloeckaert. *PLOS ONE*
462 15 (6): e0234636. https://doi.org/10.1371/journal.pone.0234636.
463 Keen, Eric C. 2015. "A Century of Phage Research: Bacteriophages and the Shaping of Modern
464 Biology: Cause to Reflect." *BioEssays* 37 (1): 6–9.
465 https://doi.org/10.1002/bies.201400152.
466 Klyczek, Karen K., J. Alfred Bonilla, Deborah Jacobs-Sera, Tamarah L. Adair, Patricia Afram,
467 Katherine G. Allen, Megan L. Archambault, et al. 2017. "Tales of Diversity: Genomic and
468 Morphological Characteristics of Forty-Six Arthrobacter Phages." Edited by Raymond
469 Schuch. *PLOS ONE* 12 (7): e0180517. https://doi.org/10.1371/journal.pone.0180517.
470 Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. "Predicting
471 Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete
472 Genomes." *Journal of Molecular Biology* 305 (3): 567–80.
473 https://doi.org/10.1006/jmbi.2000.4315.
474 Krumsiek, J., R. Arnold, and T. Rattei. 2007. "Gepard: A Rapid and Sensitive Tool for Creating
475 Dotplots on Genome Scale." *Bioinformatics* 23 (8): 1026–28.
476 https://doi.org/10.1093/bioinformatics/btm039.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 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.

477 Marchler-Bauer, Aron, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz,
478 Lewis Y. Geer, Renata C. Geer, et al. 2014. “CDD: NCBI’s Conserved Domain
479 Database.” *Nucleic Acids Research*, gku1221.
480 Mavrich, Travis N., and Graham F. Hatfull. 2017. “Bacteriophage Evolution Differs by Host,
481 Lifestyle and Genome.” *Nature Microbiology* 2 (9): 17112.
482 https://doi.org/10.1038/nmicrobiol.2017.112.
483 Melo, Anna Cristhina Carmine de, Amanda da Mata Gomes, Fernando L. Melo, Daniel M. P.
484 Ardisson-Araújo, Agueda Palmira Castagna de Vargas, Valessa Lunkes Ely, Elliot W.
485 Kitajima, Bergmann M. Ribeiro, and José Luiz Caldas Wolff. 2019. “Characterization of a
486 Bacteriophage with Broad Host Range against Strains of Pseudomonas Aeruginosa
487 Isolated from Domestic Animals.” *BMC Microbiology* 19 (1): 134.
488 https://doi.org/10.1186/s12866-019-1481-z.
489 Miller-Ensminger, Taylor, Andrea Garretto, Jonathon Brenner, Krystal Thomas-White, Adriano
490 Zambom, Alan J. Wolfe, and Catherine Putonti. 2018. “Bacteriophages of the Urinary
491 Microbiome.” *Journal of Bacteriology* 200 (7). https://doi.org/10.1128/JB.00738-17.
492 Ngandu, Nobubelo K, Konrad Scheffler, Penny Moore, Zenda Woodman, Darren Martin, and
493 Cathal Seoighe. 2008. “Extensive Purifying Selection Acting on Synonymous Sites in
494 HIV-1 Group M Sequences.” *Virology Journal* 5 (1): 160. https://doi.org/10.1186/1743-
495 422X-5-160.
496 Pedulla, Marisa L., Michael E. Ford, Jennifer M. Houtz, Tharun Karthikeyan, Curtis Wadsworth,
497 John A. Lewis, Debbie Jacobs-Sera, et al. 2003. “Origins of Highly Mosaic
498 Mycobacteriophage Genomes.” *Cell* 113 (2): 171–82. https://doi.org/10.1016/s0092-
499 8674(03)00233-2.
500 Pope, Welkin H, Charles A Bowman, Daniel A Russell, Deborah Jacobs-Sera, David J Asai,
501 Steven G Cresawn, William R Jacobs Jr, et al. 2015. “Whole Genome Comparison of a
502 Large Collection of Mycobacteriophages Reveals a Continuum of Phage Genetic
503 Diversity.” Edited by Roberto Kolter. *ELife* 4 (April): e06416.
504 https://doi.org/10.7554/eLife.06416.
505 Pope, Welkin H., and Deborah Jacobs-Sera. 2018. “Annotation of Bacteriophage Genome
506 Sequences Using DNA Master: An Overview.” In *Bacteriophages*, edited by Martha R.J.
507 Clokie, Andrew M. Kropinski, and Rob Lavigne, 1681:217–29. Methods in Molecular
508 Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7343-
509 9_16.
510 Pope, Welkin H., Travis N. Mavrich, Rebecca A. Garlena, Carlos A. Guerrero-Bustamante,
511 Deborah Jacobs-Sera, Matthew T. Montgomery, Daniel A. Russell, Marcie H. Warner,
512 Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary
513 Science (SEA-PHAGES), and Graham F. Hatfull. 2017. “Bacteriophages of Gordonia
514 Spp. Display a Spectrum of Diversity and Genetic Relationships.” *MBio* 8 (4).
515 https://doi.org/10.1128/mBio.01069-17.
516 Russell, Daniel A. 2018. “Sequencing, Assembling, and Finishing Complete Bacteriophage
517 Genomes.” In *Bacteriophages: Methods and Protocols, Volume 3*, edited by Martha R.J.
518 Clokie, Andrew M. Kropinski, and Rob Lavigne, 109–25. Methods in Molecular Biology.
519 New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-7343-9_9.
520 Russell, Daniel A., and Graham F. Hatfull. 2017. “PhagesDB: The Actinobacteriophage
521 Database.” *Bioinformatics (Oxford, England)* 33 (5): 784–86.
522 https://doi.org/10.1093/bioinformatics/btw711.
523 Sahoo, Dinabandhu, N. Jusna Devi, N. Ngashangva, P. Momota, Y. Rojeena, and S. Indira
524 Devi. 2019. “Draft Genome Sequence of Arthrobacter Globiformis Mrc11, an
525 Antimicrobial Agent Isolated from a Khangkhui Cave Deposit.” Edited by Catherine
526 Putonti. *Microbiology Resource Announcements* 8 (11): MRA.01620-18, e01620-18.
527 https://doi.org/10.1128/MRA.01620-18.

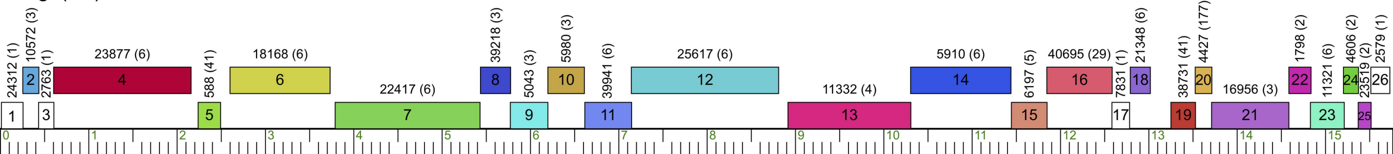
bioRxiv preprint doi: https://doi.org/10.1101/2020.08.15.252577; this version posted August 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

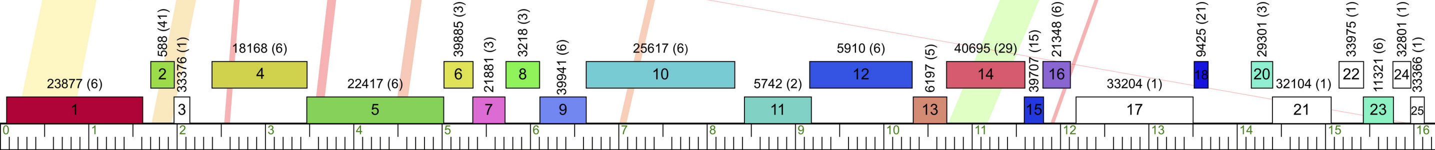
528 Schneider, Caroline A, Wayne S Rasband, and Kevin W Eliceiri. 2012. "NIH Image to ImageJ:
529 25 Years of Image Analysis." *Nature Methods* 9 (7): 671–75.
530 https://doi.org/10.1038/nmeth.2089.
531 *SEA-PHAGES/Starterator*. (2016) 2020. Python. SEA-PHAGES Program.
532 https://github.com/SEA-PHAGES/starterator.
533 Shi, Yibo, Ning Li, Yaxian Yan, Hengan Wang, Yan Li, Chengping Lu, and Jianhe Sun. 2012.
534 "Combined Antibacterial Activity of Phage Lytic Proteins Holin and Lysin from
535 Streptococcus Suis Bacteriophage SMP." *Current Microbiology* 65 (1): 28–34.
536 https://doi.org/10.1007/s00284-012-0119-2.
537 Soding, J., A. Biegert, and A. N. Lupas. 2005. "The HHpred Interactive Server for Protein
538 Homology Detection and Structure Prediction." *Nucleic Acids Research* 33 (Web
539 Server): W244–48. https://doi.org/10.1093/nar/gki408.
540 Tsirigos, Konstantinos D., Christoph Peters, Nanjiang Shu, Lukas Käll, and Arne Elofsson.
541 2015. "The TOPCONS Web Server for Consensus Prediction of Membrane Protein
542 Topology and Signal Peptides." *Nucleic Acids Research* 43 (W1): W401-407.
543 https://doi.org/10.1093/nar/gkv485.
544 Wiles, Siouxsie. 2014. "Monday Micro – 200 Million Light Years of Viruses?!" Sciblogs. May 8,
545 2014. https://sciblogs.co.nz/infectious-thoughts/2014/08/05/monday-micro-200-million-
546 light-years-of-viruses/.
547 Yuan, Yihui, and Meiying Gao. 2017. "Jumbo Bacteriophages: An Overview." *Frontiers in
548 Microbiology* 8 (March). https://doi.org/10.3389/fmicb.2017.00403.
549
550

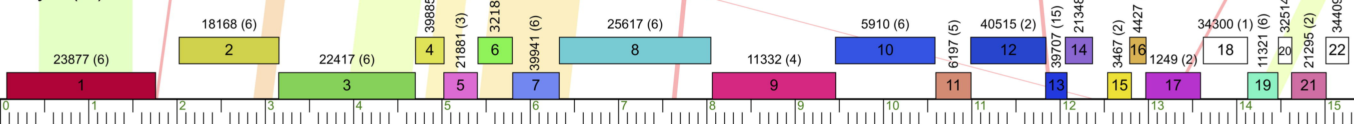
A

B

20 nm


Corgi (FE)


BlueFeather (Singleton)


Whytu (FI)

A

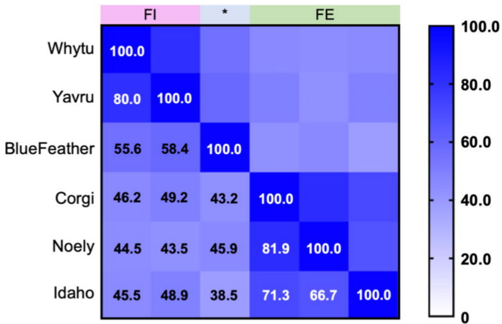
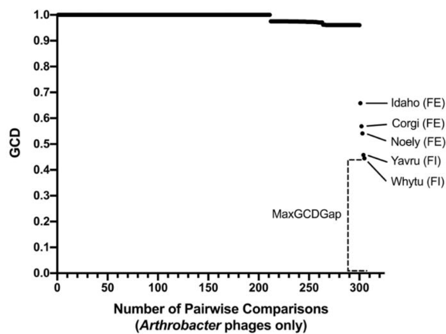
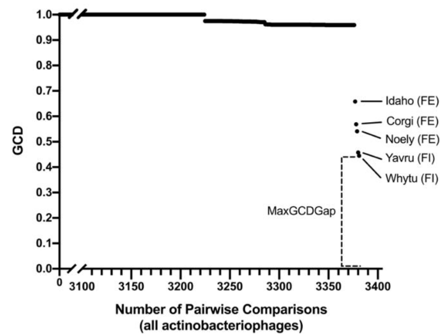
FI * FE

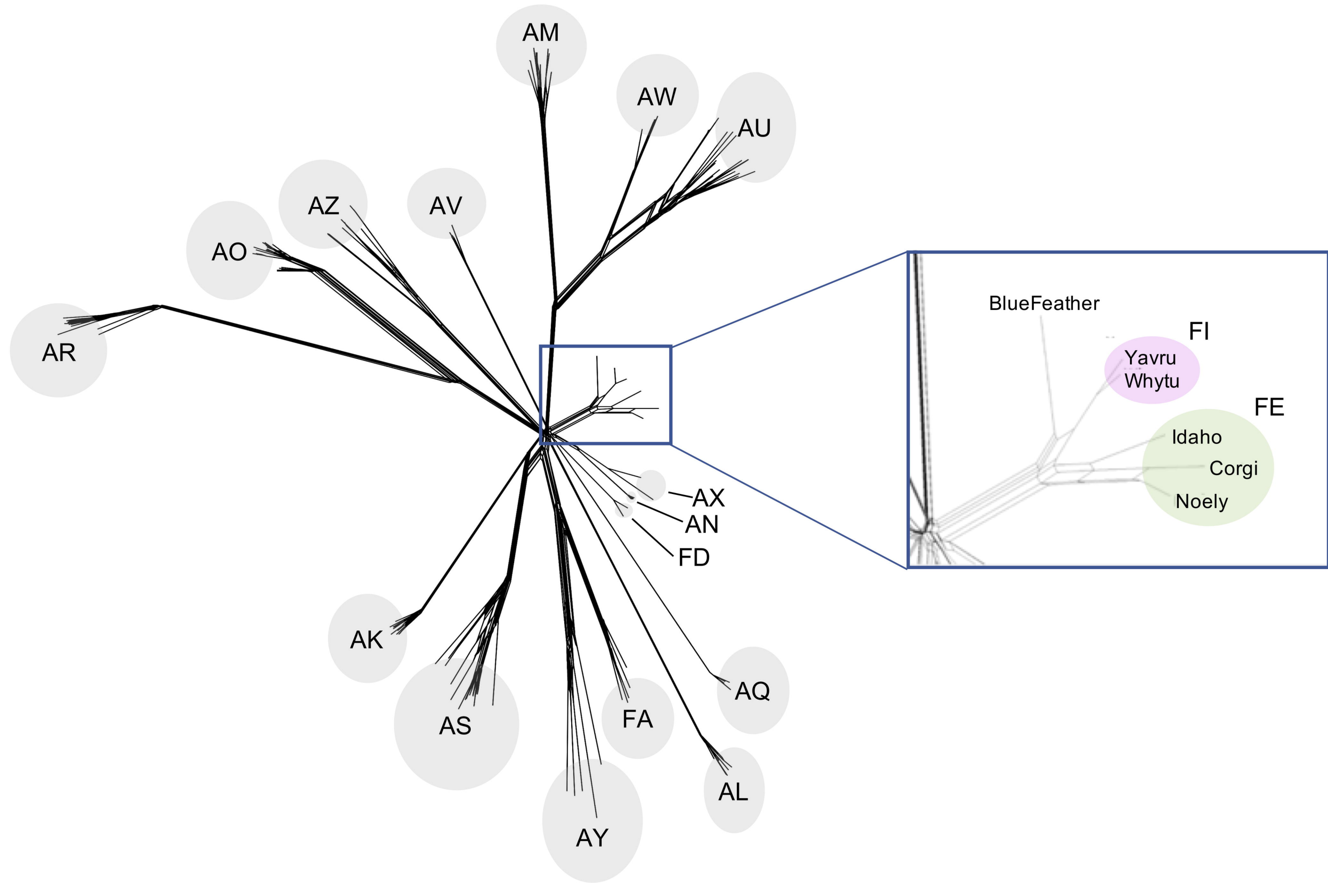
Yavru
Whytu
BlueFeather
Corgi
Noely
Idaho

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
AQ
AL
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