

Comparison of Mycobacteriophage Cluster K Tail Proteins in Relation to Host Range

Gabrielle Shapiro

Abstract

Mycobacteriophages are viruses that are parasitic to the fungus-like *Mycobacterium* and with the recent increase in antibiotic resistance in these bacteria such as *M. tuberculosis*, more research into using phages as a treatment is imperative. A larger host-range, which has been shown to be connected to tail proteins, is preferred when searching for the specific phage to use in the treatment “cocktails” because it vastly reduces the amount of time it takes to find the right phage. However, testing for host-range has many challenges, from plating dangerous bacteria to time constraints, thus it is crucial for there to be a way to test host range without the need for in-person plaque tests. I used several online technologies: first, the fully sequenced phage genomes are found on Actinobacteriophage Database (PhagesDB.org), then the Phamerator compares genes between different phages. BLASTP (P for protein) finds related sequences. $1e-20$ will be used as a cut-off value, defining a “strong” relationship. The resulting hits were compared with Cluster data in previous research including host range. Using MUSCLE on MEGAX, sequences underwent multiple sequence alignment (.FASTA sequences from Batch Entrez). Neighbor-Joining phylogenetic trees were built for each tail protein BLAST and one for the tape measure protein of the same phages. A Mann-Whitney U test was used to evaluate whether the sum of branch lengths (evolutionary distance) are statistically significant. I investigated how host range can be determined on a genetic basis. Indeed, Cluster K host range is known to be large and connected to its tail proteins, thus I compared them to tail proteins in other

mycobacteriophages. I expected to see statistically significant results on phages that also have larger host ranges. I saw correlations in certain phages to larger host ranges.

Introduction

Bacteriophages are viruses that infect and lyse bacteria. Phages are reported to be some of the most diverse, abundant (McNerney, R., & Traoré, 2005), and oldest (Abrescia *et al*, 2012) organisms on earth. Initially described a century ago by William Twort (Twort, 1915) and Felix d'Herelle (Bruynoghe, 1921), phages have been a pivotal tool in the study of virology, genetics, and human health. Once antibiotics were discovered as a much simpler and cross-species way to kill bacteria, phage research was mostly pushed aside. However today, multi-drug resistance is a growing trait in bacteria, re-emphasizing the necessity of research into bacteriophages (Potera C, 2013).

Mycobacterium is a genus of Actinobacteria, named due to their fungus-like behavior. Many of the genomes of the >190 species of *Mycobacterium* are now fully sequenced. Thus *Mycobacterium* and *mycobacteriophages* serve as a crucial tool in the study of viral diversity and evolution. The *Mycobacterium* genus also includes several pathogens that are known to cause serious diseases in humans, including *M. tuberculosis* (TB) and *M. leprosy*. TB continues to be one of the deadliest diseases globally, killing over 1.5 million people a year (Glaziou *et al*, 2014). There is also a continued concern from the evolution of drug resistance in TB (Gillespie S. H., 2002). Therefore, it is imperative to continue researching mycobacteriophages as a treatment for TB.

Unfortunately, *M. tuberculosis* is not an ideal host to use in isolating phages, due both to its dangerous nature and its remarkably slow growth rate. Instead,

Mycobacterium smegmatis, a fast-growing, and non-pathogenic organism, is used as the model host. The first mycobacteriophage was isolated from *M. spegmitis* in the 1940s (Gardner Gm, Weiser Rs, 1947). In the early 1990s, with the onset of new sequencing and gene editing technologies, the first complete mycobacteriophage genome was published (Hatfull, G. F., & Sarkis, G. J, 1993).

Individual phages are isolated from a simple plaque assay and can be found in many natural environments from soil to compost samples. The Howard Hughes Medical Institute (HHMI) and Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) have created a course for undergraduate students that is widely implemented and taken by over 4,800 students at 73 institutions (Jordan, T. C *et al*, 2014). Currently, this work has allowed for the discovery of over 18,000 total found phages, with around 2,000 sequenced mycobacteriophages.

The current knowledge of mycobacteriophage genomes shows that they have been actively engaged in horizontal genetic exchange for several billion years, leading to one of their most striking features, pervasive mosaicism. The lineage of certain related genes, which are grouped into phamilies (phams), come from many different sources. A likely major contributor to the mosaicism is illegitimate recombination with natural selection for function (Hatfull, G. F., & Hendrix, R. W, 2011). Nonetheless, phages can be organized by their nucleotide sequence similarity. If phages share over 35% of their genome (because there is a noticeable gap there in gene content similarity across all phages), then they are put together in a cluster. Clusters are designated A-Z and more distantly related phages are called singletons. There is also a correlation between cluster and host range (Jacobs-Sera, D. *et al*, 2012). In addition, since Tape Measure Proteins are highly conserved, they can also be used to define clusters (Smith et al, 2013). For

example, Cluster K is correlated strongly with larger host ranges and current research suggests that certain tail proteins contribute to this broader host range (Dedrick RM *et al*, 2019a).

Host range describes the number of host species a parasite is capable of infecting and in the case of bacteriophages it is the taxonomic diversity of hosts the phage can successfully infect (Jonge, P. A *et al*, 2018). For Bacteriophage therapy, a broad host range is consistently preferred because a narrow range limits the phage's use to a small set of pathogens. (Ross, A. *et al*, 2016). For example, phage host range definitions were used to narrow down 10,000 phages to a specific cocktail that succeeded in curing a patient infected with *M. abcessus* (Dedrick RM *et al*, 2019b). Using the large host range exponentially reduced the amount of time it took to find the correct phage. This use of host range is tremendously important as it takes an extremely long time to search for and test phages; which has led to patient deaths. However, testing for host range has many challenges, from plating dangerous bacteria to time constraints. For this reason, finding a way to test host range without the need for in-person plaque tests is crucial.

There is a strong connection between host range and tail proteins. Minor tail proteins are some of the most diverse sequences that contribute to adsorption, penetration, and importantly recognition (Hatfull G. F., 2014a). In other genera of bacteria, such as *Staphylococcus aureus*, the discovery of tail proteins allowing for a large host range allowed scientists to create models predicting host range based solely on genetic material (Moller, A. G *et al*, 2020). In *Listeria monocytogenes*, comparative genomics was used to find receptor binding proteins that were crucial for host specificity (Casey, A. *et al*, 2015). Then these same proteins are reprogramed in other phages to increase host range (Dunne, M. *et al*, 2019). This correlation is also seen in

mycobacteriophages where mutations on tail protein 22 allowed for phages to infect different hosts (Jacobs-Sera, D. *et al*, 2012). Bioinformatic approaches exist to compare proteins in different mycobacteriophage clusters for host range (Norian, H. A, 2014-2015).

I investigated whether host range can be determined on a genetic basis. Indeed, Cluster K host range is known to be large and connected to its tail proteins (Pope WH, *et al*, 2011), thus I compared them to tail proteins in other mycobacteriophages. Previous research has also outlined a larger host range in some A1, A2, A3, A9, B1, B2, G, and K phages (Jacobs-Sera, D. *et al*, 2012), (Hatfull G. F, 2014b), as such, I expected to see related tail proteins as indicated by more BLAST hits on these clusters.

Results

I chose ten phages: Angelica, CrimD, TM4, ZoeJ, Pixie, Fionnbharth, Larva, Guillsminger, Krueger, Aminay, one to two per each of the seven clusters. These phages are phages that have already been shown to have a large host range so that there are previous results to compare to. There were 8-12 tail proteins per phage, several of them were from the same pham, such as 46982: the same pham was used by Jacobs-Sera to increase host range. However many other phams were unique to each phage. In addition, the number of codons varied greatly between tail proteins with the largest usually tail protein of 19 or 23. Refer to Fig 1 for the Phamerator of the tail proteins in these phages.

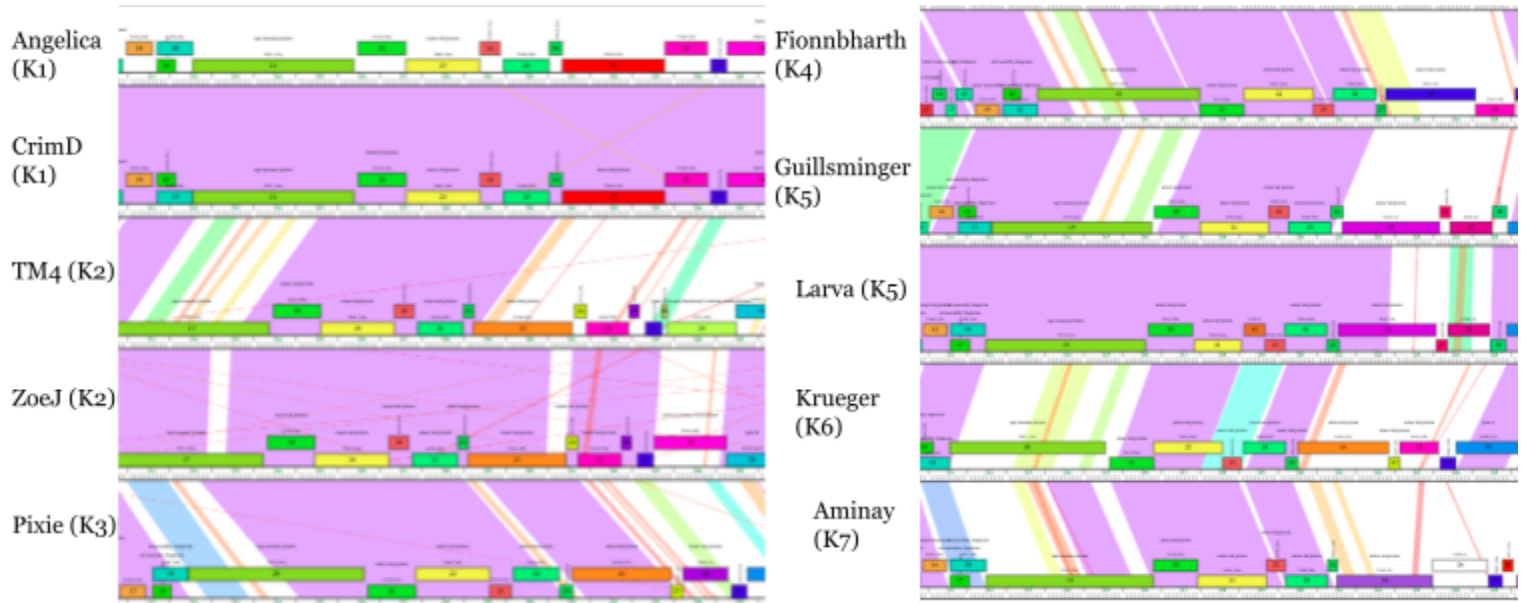


Fig. 1 Phamerator for chosen phages. The shading is color-coded according to the E-value with violet representing an E value of zero and red the threshold value.

I used BLAST (P) on each of these tail proteins. The two K2 phages are used to discuss BLAST (P) hits. Refer to Supplementary Material 1 to view the results from BLAST hits. Cluster K2 BLASTP hits included: A1, A10, A11, A12, A13, A14, A15, A17, A18, A2, A20, A3, A4, A5, A6, A7, A9, AB, AC, B1, B10, B11, B3, B7, B8, B9, CD, ChewyVIII, CQ1, CQ2, CR1, CR3, CR4, CR5, CS1, CS3, CT, CV, DA, DC1, DC2, DE1, DE2, DE3, DE4, DE5, DI, DK, DL, DP, DS, DS6A, DT, DW, DY, E, EP, F1, F2, F3, G1, G2, G3, G4, G5, H2, J, K1, K2, K3, K4, K5, K6, K7, L3, L4, LilSpotty, MalagasyRose, N, P1, Q, REQ3, Sleepyhead, TPA4, Whack, X, Y, Yvonnestic, Z. Refer to Fig 2a and 2b. to see how clusters are designated.

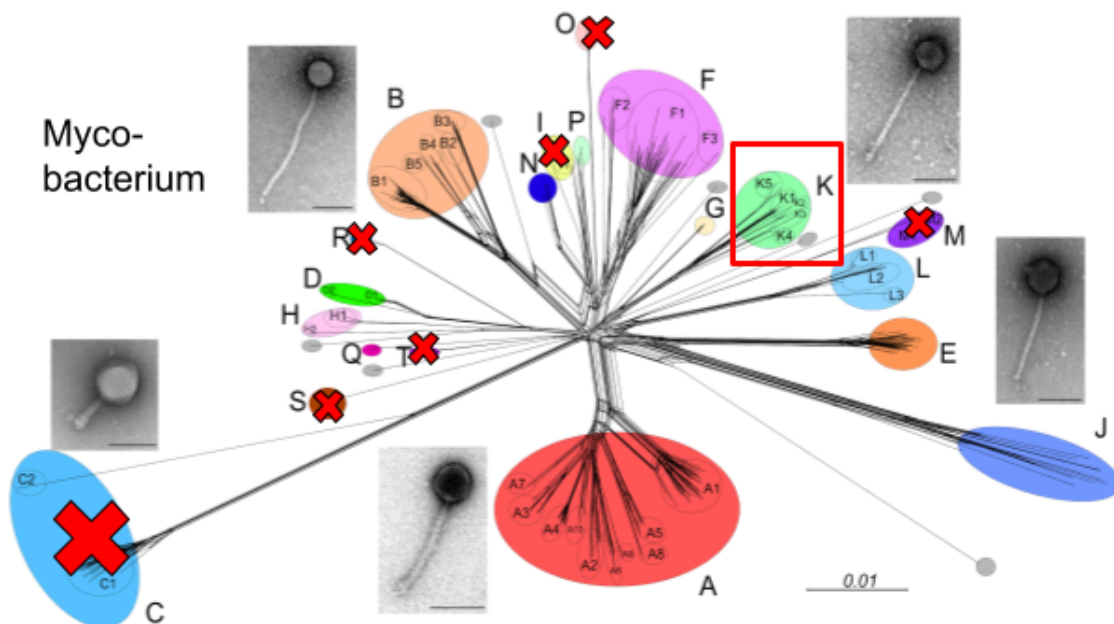


Fig 2a. Splistree network for Mycobacteriophages. X's indicate no hits and there is a square to point out cluster K. From Hatfull GF (2014) Mycobacteriophages: Windows into Tuberculosis. PLoS Pathog 10(3): e1003953.

<https://doi.org/10.1371/journal.ppat.1003953>.

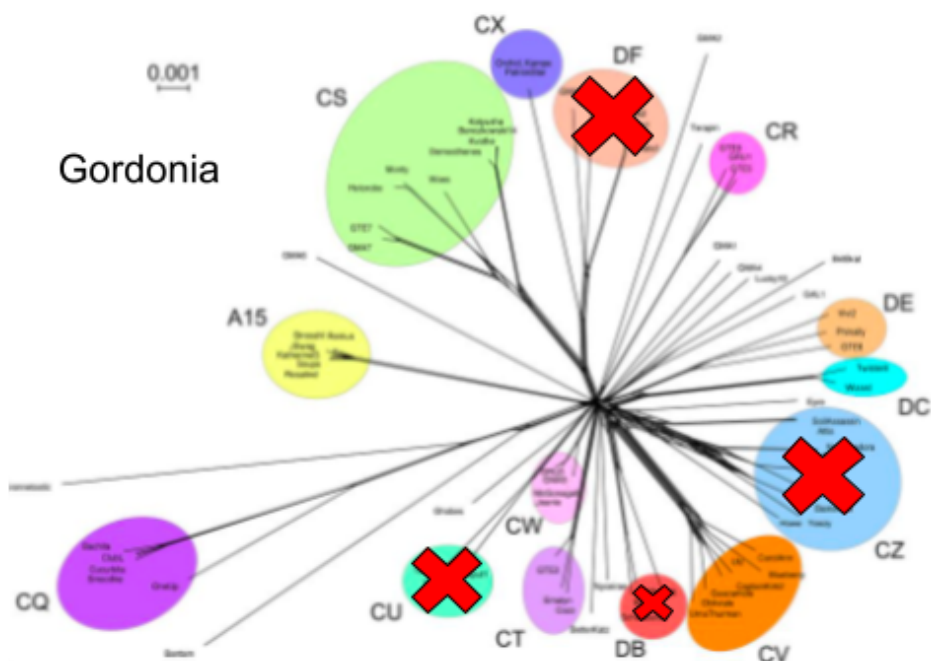


Fig 2b. Splitstree Network of Gordonia phages. From

Welkin H. Pope, Travis N. Mavrich, Rebecca A. Garlena, Carlos A. Guerrero-Bustamante, Deborah Jacobs-Sera, Matthew T. Montgomery, Daniel A. Russell, Marcie H. Warner, Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES), Graham F. Hatfull, mBio Aug 2017, 8 (4) e01069-17; DOI: 10.1128/mBio.01069-17

In addition Fig 3. portrays the Blast hits from TM4 and ZoeJ respectively. The most closely related phages (smallest E-value) for each protein was from the subcluster that the first tail protein came from, next there was the cluster. These data show the differences between cluster designation and how closely related the tail proteins are. For example, there are many clusters that are more distantly related but have a smaller E-value. Cluster L3 is more distantly related from K2, and yet its hits show that their tail proteins are very similar. This denotation is important because host-range isn't based on the whole genome and as such simple cluster designations cannot fully describe host range. Cluster Ks and A2 and A3 are not closely related and yet they both are shown to have larger host ranges. Another piece of information to look at is the number of tail proteins that are similar between phages. As expected, almost every tail protein from cluster K2 has a hit from the other Cluster Ks. Of the Cluster A subclusters, A2 and A3 have the most number, 3, of significantly related tail proteins. There is a clear upward trend, which can be explained because when a cluster is more related to K2, its tail proteins are also related. However, this trend is not true for every cluster. Drops indicate a closely related tail protein from a more distantly related phage.

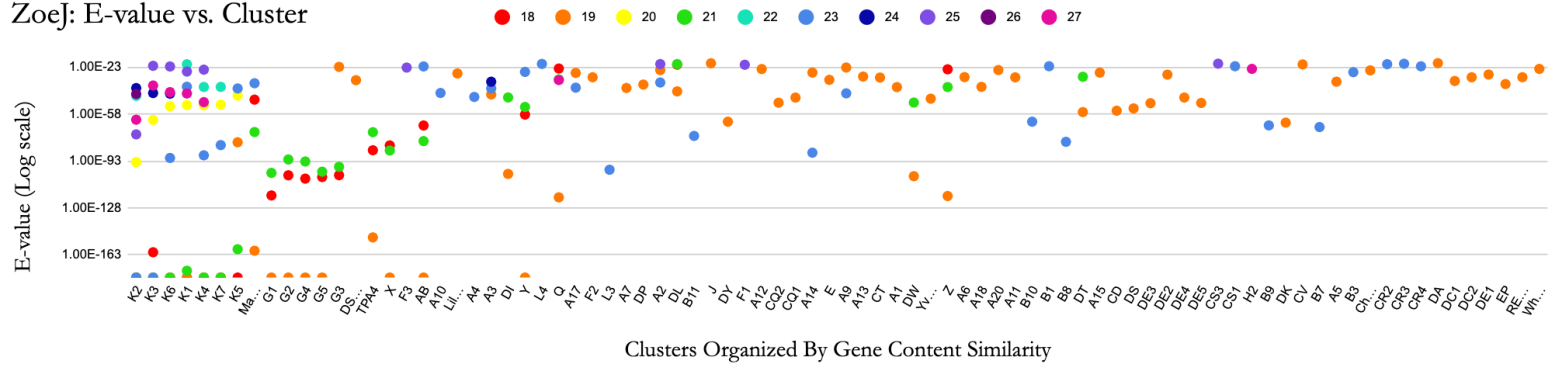
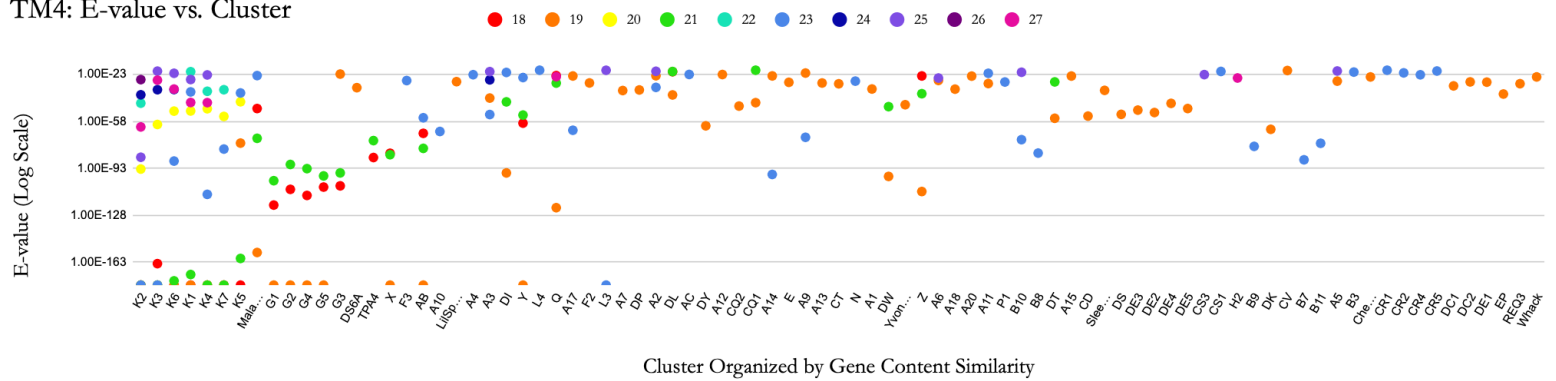


Fig 3. For each pham (gene phamily) the BLAST hits are averaged per cluster and graphed against the other phams. When the value is smaller, it indicates a closer related tail protein. The legend describes proteins 18-27. E-value decreases exponentially when matches are closer, thus a log scale helps show this trend. Smaller E-Values indicate closer relation. E-value is the expected hits to see by chance when searching the large (~100GB) BLAST database. There is a clear upward trend followed GCS (cluster designation).

While these results show clear trends and correlations between different clusters and tail protein similarity, it cannot be shown to be statistically significant as there aren't any E-values from the whole genome to compare to. Thus phylogenetic trees were used to test the statistical significance of these correlations. Phylogenetic trees based on TM4 and the tape measure protein were created. Fig 4. shows the neighbor-joining

phylogenetic trees for TM4. When each gene is compared to the tape measure tree on Phylo.io, there are many similar nodes, using a variant on the Jaccard index.

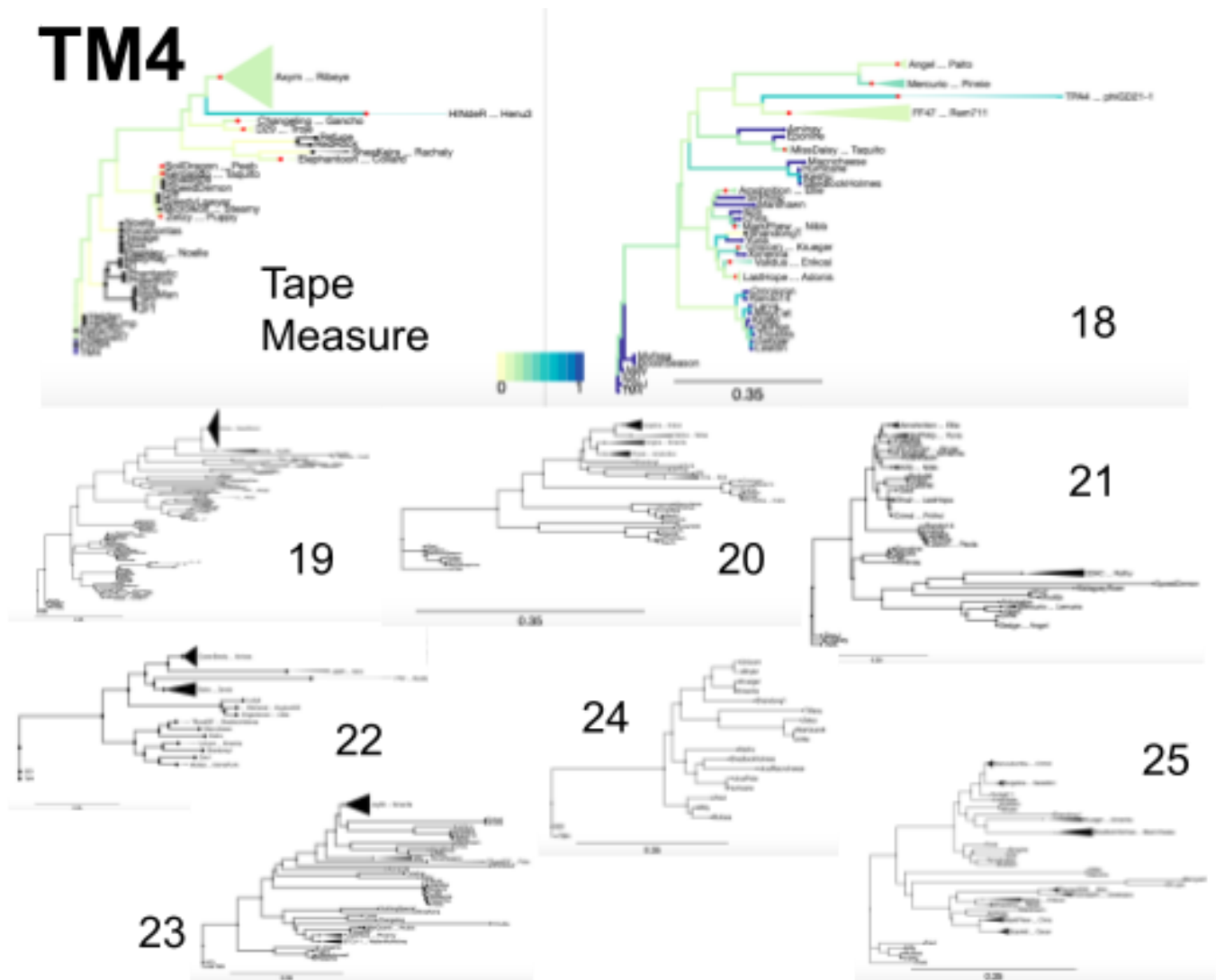


Fig 4. Neighbor Joining phylogenetic trees for TM4: Tape Measure protein and minor tail proteins 18-25. Scale 0.35 substitutions per site. Some trees are larger or smaller and this depends on the number of blast hits along with the sequence lengths. Each tree rooted on TM4 to allow for a basis of comparison. Top trees have been compared while the bottom are solely portrayed.

After I parsed all the branch lengths with the code on google sheets, I was able to statistically test them. Supplementary Material 2, gives a list of all branch lengths. Fig 5. shows a graph of these branch lengths for TM4 and the Tape Measure Protein trees. As expected branch lengths were smallest on the closely related phages such as cluster K, but more distantly related phages also had larger branch lengths. In addition the tape measure protein lengths were larger for most clusters except a few. This shows how the cluster designation is further apart than just the tail proteins.

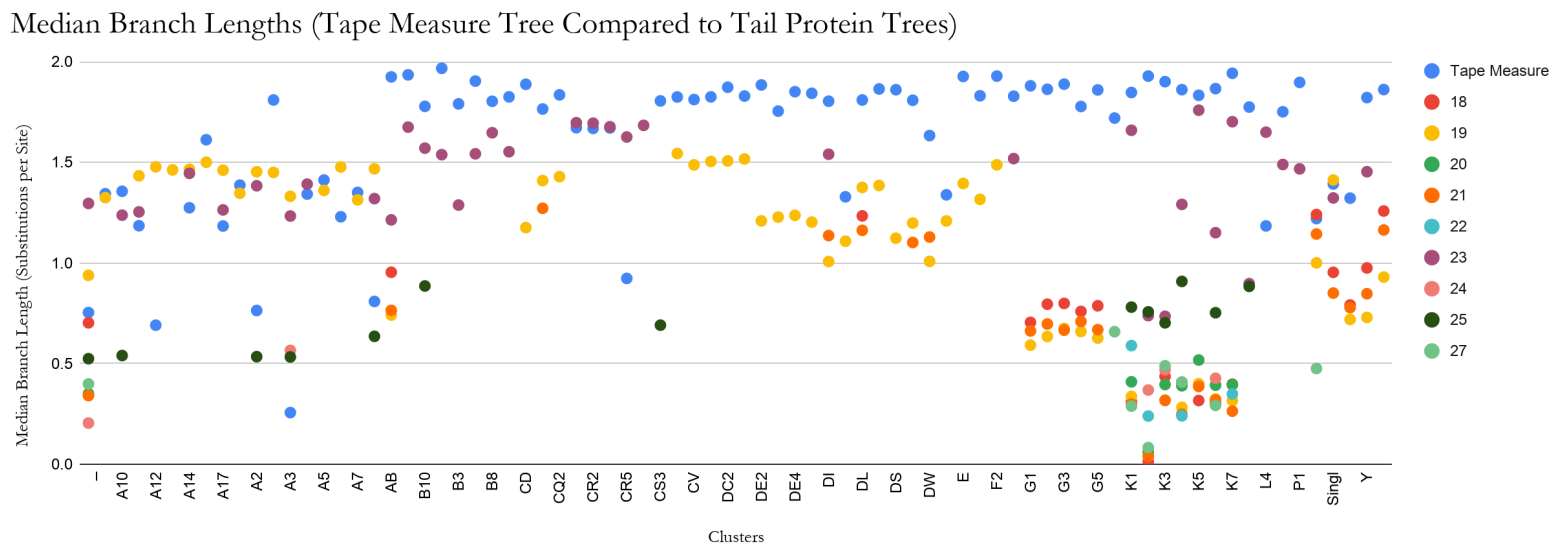


Fig 5. Median Branch Lengths. Y axis describes median branch lengths in substitutions per site, while the X axis shows each cluster. The legend is tail proteins 18-27.

Next a Mann Whitney U-Test was performed. This allowed for the results to be judged on how significant they were. The assumptions were fulfilled: I had one dependent variable (branch length) that was measured continuously, one independent variable that consisted of two categorical, independent groups (Tape Measure phylogenetic tree vs Minor tail protein tree), and the groups were independent. This test

was specifically non-parametric, two samples, and heteroscedastic. My null hypothesis (Ho) was that the tail protein sum of branch lengths and the tape measure protein sum of branch lengths have the same distribution of scores. My research hypothesis (H) was that the tail proteins are more closely related than just cluster designation, thus the tail protein sum of branch lengths is smaller than the tape measure protein sum of branch lengths. Refer to Fig 6. to see the results per cluster.

	Supports Research H	Does Not Support Research H
Reject Null H ₀	A1 A10 AB B3 CD CS3 CT CV DC1 DE1 DI DL E G1 G2 G4 K1 K2 K3 K4 K5 K6 L3 Q Y (average common language effect size 87%)	A6 A9 A11 CR2 CR4 A2 A3 A4
Cannot Reject H ₀	A18 A20 AC B7 B8 B9 B10 B11 CQ1 CQ2 DC2 DE2 DE3 DE4 DE5 DK DP DS DT DW DY EP F2 F3 G3 H2 K7 N P1 X Z	A12 A13 A14 A15 A17 CR1 CR5 CS1 G5 L4

Fig 6. Chart of clusters that could reject Ho and support H.

Of the clusters that can reject the null hypothesis and support research hypothesis: on average the medians for the tail protein branch lengths were 0.84 substitutions per site smaller, with common language effect size ranging from 63% to 100% (mean 87%). Average R^2 of 0.76. 87% of the rankings were favorable to the

research hypothesis. This effect size shows how the difference between branch lengths is large. Effect size is less affected by sample size, which is important as sample sizes vary greatly.

Discussion

My results show that there is a clear relationship between the tail proteins and host range, however, this relationship depends on the cluster. Some tail proteins are more closely related than gene content similarity (or the alternative ways to define clusters). Many clusters did not have any BLAST hits, indicating a smaller host range. There were many hits on *Gordonia* phages, indicating a tail protein relationship and larger host ranges.

Comparing the expected results, reveals some interesting points about different clusters. A1, A2, A3, A9, B1, G, and K with larger host ranges were seen. This shows that even though they are more distantly related when just looking at cluster designation, their tail proteins were significantly related. Because I saw the hits that were expected, that means that the other phages with significant results are correlated with larger host ranges. Unfortunately, there are so many phages out there that the vast majority have never been tested for host range, however these results indicate which ones have an expected larger range. B2 was not a hit, B1 was not a hit for TM4 but, for example, ZoeJ had many B1 hits. However, A2, A3, A9 tail proteins were more distantly related than if they were just compared by cluster. These cannot be disregarded as they still have related tail proteins that are statistically significant; simply, there may be something with more impact on host range than tail proteins in these phages.

Next, after the Mann Whitney U-test, only some phages' tail proteins were statistically significantly more related than if they were just to be compared to cluster designation. These phages supported research hypotheses about tail proteins being more closely related than cluster designation. Interestingly, some tail proteins were more distantly related.

These results show new relations between different cluster tail proteins. The phages that were shown statistically significant are correlated with larger host range (A1, A10, AB, B3, CD, CS3, CT, CV, DC1, DE1, DI, DL, E, G1, G2, G4, K1, K2, K3, K4, K5, K6, L3, Q, Y). In addition they support the previous findings around tail proteins and host range determinations.

However, there are a few limitations to this research approach. Since E-value is proportional to sequence size, this may affect the branch lengths on phylogenetic trees influencing the smaller trees to be wrongly calculated as statistically significant. The MUSCLE multiple sequence alignment or BLAST could have been set to the wrong parameters. Mann-Whitney U tests are more likely to reject a true null hypothesis when data are both heteroscedastic and non-normal.

There are many places for further research. The other 9 Cluster K phages need to be statistically tested. These new tests may show other relations, and help with the issue of some clusters having to reject null hypotheses as there were just too few phages to gain any statistical significance. Many of the BLAST hits were on bacterium, leading to the question, if host range can be determined on types of bacteria with which they have undergone horizontal transfer.

Applications for these results are numerous. First, this research helps speed up the time it takes to find the correct cocktail. While individually there is no way to state-

because a certain phage has a closely related tail protein they must also have a large host range- what this research shows is the correlation. Scientists have to test tens of thousands of phages and knowing which phages are correlated to a larger host range will save extraordinary amounts of time. These results help indicate which phages to test for host range. This process for defining host range can be used on other viruses, from other phages to zoonotic viruses. For example, it is crucial to know how viruses such as SARS-CoV-2 cross species and the mechanisms to do so. Discovering correlations in host range and tail proteins is critical and my research broadens this understanding.

Methods

I used several online technologies to compare Cluster K phage tail proteins with tail proteins from other phages. A stronger relationship between the tail proteins implied that the other phages also have a broader host range. The fully sequenced genomes were found on Actinobacteriophage Database (PhagesDB.org) a comprehensive, interactive, database-backed website (Russell, D. A., & Hatfull, G. F, 2017). This database describes the clusters and changes phage designation with new research. Additionally, it hosts an API that I used to fill in cluster information. To compare the genes between different phages, the Phamerator is used. The Phamerator creates a visual representation of the protein-coding genes into phamilies of related sequences, which is crucial due to the mosaicism of bacteriophages (Cresawn, S.G *et al*, 2011). Lastly, I used Basic Local Alignment Search Tool (BLAST) or specifically BLASTP (P for protein) to search for related tail-proteins in other phages. Here BLASTP is used because the amino acids differ far less than the actual nucleotides. BLAST or Basic Local Alignment Search Tool is crucial to find phages with resemblance between biological

sequences (Altschul, S. F *et al*, 1990). An E-value of $1e-20$ is used to define a “strong” relationship. The resulting hits were compared with cluster data in previous research such as host range. Graphs were organized by gene content similarity (GCS), a metric to measure proportion of shared genes. This organization showed differences in expected cluster designation and the resulting tail protein relatedness.

Neighbor Joining (N Saitou, M Nei, 1987) phylogenetic trees based on the sequence results from BLAST were built using MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms (Kumar *et al*, 2018) for Mac OS (Stecher *et al*, 2020). An additional tree was built from the tape measure proteins of these same phages. The tape measure proteins are highly conserved and can be used as a simpler method to define clusters (Smith *et al*, 2013). The .FASTA sequences were downloaded by inputting accession numbers into NCBI Batch Entrez and multiple sequence global alignment was performed using MUSCLE (Edgar R. C., 2004) with default gap penalties and UPGMA (unweighted pair group method with arithmetic mean) cluster method. Phylo.io was used to visualize the resulting phylogenetic trees and to root the tree on the Cluster K phage used in the original BLAST (Oscar, 2016). The average sum of branch lengths (evolutionary distance) was calculated for each cluster type from the phylogenetic tree through a code in google sheets. The non-parametric, two sample, heteroscedastic Mann-Whitney U test was used to compare the average branch lengths per cluster for the tail protein trees and the tape measure tree; a p value of 0.05 will be used as the cutoff. The common language effect size was calculated. Finally, to mitigate any risks from Covid-19, all research was conducted on my home computer and any meetings with scientists were online.

Links to Online Software

Actinobacteriophage database: <https://phagesdb.org/>

Phamerator: <https://phamerator.org/>

BLAST: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Batch Entrez: <https://www.ncbi.nlm.nih.gov/sites/batchentrez>

MEGAX: <https://www.megasoftware.net/>

Phylo.io: <http://phylo.io/index.html#>

Literature Cited

Abrescia, N. G., Bamford, D. H., Grimes, J. M., & Stuart, D. I. (2012). Structure Unifies the Viral Universe. *Annual review of biochemistry*, 81, 795–822.

<https://doi.org/10.1146/annurev-biochem-060910-095130>

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), 403-410.

<https://pubmed.ncbi.nlm.nih.gov/2231712/>

Bruynoghe, R. A. J. M., & Maisin, J. (1921). Essais de Thérapeutique au Moyen du Bacteriophage. *CR Soc Biol*, 85, 1120-1121.

Casey, A., Jordan, K., Neve, H., Coffey, A., & Mcauliffe, O. (2015). A Tail of Two Phages: Genomic and Functional Analysis of *Listeria monocytogenes* Phages vB_LmoS_188 and vB_LmoS_293 Reveal the Receptor-Binding Proteins Involved in Host Specificity. *Frontiers in Microbiology*, 6, 1107.

[doi:10.3389/fmicb.2015.01107](https://doi.org/10.3389/fmicb.2015.01107)

- Cresawn, S.G., Bogel, M., Day, N. et al. Phamerator: a Bioinformatic Tool for Comparative Bacteriophage Genomics. BMC Bioinformatics 12, 395 (2011).
<https://doi.org/10.1186/1471-2105-12-395>
- Dedrick, R. M., Bustamante, C. A., Garlena, R. A., Pinches, R. S., Cornely, K., & Hatfull, G. F. (2019). Mycobacteriophage ZoeJ: A Broad Host-Range Close Relative of Mycobacteriophage TM4. Tuberculosis, 115, 14-23.
[doi:10.1016/j.tube.2019.01.002](https://doi.org/10.1016/j.tube.2019.01.002)
- Dedrick RM, Guerrero-Bustamante CA, Garlena RA, et al. (2019). Engineered Bacteriophages for Treatment of a Patient with a Disseminated Drug-Resistant *Mycobacterium abscessus*. Nat Med. 2019;25(5):730-733.
[doi:10.1038/s41591-019-0437-z](https://doi.org/10.1038/s41591-019-0437-z)
- Dunne, M., Rupf, B., Tala, M., Qabрати, X., Ernst, P., Shen, Y., . . . Kilcher, S. (2019). Reprogramming Bacteriophage Host Range through Structure-Guided Design of Chimeric Receptor Binding Proteins. Cell Reports, 29(5).
[doi:10.1016/j.celrep.2019.09.062](https://doi.org/10.1016/j.celrep.2019.09.062)
- Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792–1797.
<https://doi.org/10.1093/nar/gkh340>
- Gardner Gm, Weiser Rs. A Bacteriophage for Mycobacterium Smegmatis. Proc Soc Exp Biol Med. 1947 Oct; 66(1):205. doi: 10.3181/00379727-66-16037. PMID: 20270730.
- Gillespie S. H. (2002). Evolution of Drug Resistance in Mycobacterium Tuberculosis: Clinical and Molecular Perspective. Antimicrobial Agents and Chemotherapy, 46(2), 267–274. <https://doi.org/10.1128/aac.46.2.267-274.2002>

Glaziou, P., Sismanidis, C., Floyd, K., & Raviglione, M. (2014). Global Epidemiology of Tuberculosis. *Cold Spring Harbor Perspectives in Medicine*, 5(2), a017798.

<https://doi.org/10.1101/cshperspect.a017798>

Glen Stecher, Koichiro Tamura, and Sudhir Kumar (2020) Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Molecular Biology and Evolution*

<https://doi.org/10.1093/molbev/msz312>

Hatfull G. F. (2014a). Molecular Genetics of Mycobacteriophages. *Microbiology spectrum*, 2(2), 1–36.

Hatfull GF (2014b) Mycobacteriophages: Windows into Tuberculosis. *PLoS Pathog* 10(3): e1003953. <https://doi.org/10.1371/journal.ppat.1003953>

Hatfull, G. F., & Hendrix, R. W. (2011). Bacteriophages and Their Genomes. *Current Opinion in Virology*, 1(4), 298–303.

<https://doi.org/10.1016/j.coviro.2011.06.009>

Hatfull, G. F., & Sarkis, G. J. (1993). DNA Sequence, Structure and Gene Expression of Mycobacteriophage L5: a Phage System for Mycobacterial Genetics. *Molecular Microbiology*, 7(3), 395–405. <https://doi.org/10.1111/j.1365-2958.1993.tb01131.x>

Jacobs-Sera, D., Marinelli, L. J., Bowman, C., Broussard, G. W., Guerrero Bustamante, C., Boyle, M. M., Petrova, Z. O., Dedrick, R. M., Pope, W. H., Science Education Alliance Phage Hunters Advancing Genomics And Evolutionary Science Sea-Phages Program, Modlin, R. L., Hendrix, R. W., & Hatfull, G. F. (2012). On the Nature of Mycobacteriophage Diversity and Host Preference. *Virology*, 434(2), 187–201. <https://doi.org/10.1016/j.virol.2012.09.026>

Jonge, P. A., Nobrega, F. L., S. J., Brouns, & Dutilh, B. E. (2018, September 1).

Molecular and Evolutionary Determinants of Bacteriophage Host Range. Trends in Microbiology. <https://doi.org/10.1016/j.tim.2018.08.006>

Jordan, T. C., Burnett, S. H., Carson, S., Caruso, S. M., Clase, K., DeJong, R. J.,

Dennehy, J. J., Denver, D. R., Dunbar, D., Elgin, S. C., Findley, A. M.,

Gissendanner, C. R., Golebiewska, U. P., Guild, N., Hartzog, G. A., Grillo, W. H.,

Hollowell, G. P., Hughes, L. E., Johnson, A., King, R. A., ... Hatfull, G. F. (2014).

A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. mBio, 5(1), e01051-13.

<https://doi.org/10.1128/mBio.01051-13>

McNerney, R., & Traoré, H. (2005). Mycobacteriophage and Their Application to

Disease Control. Journal of Applied Microbiology, 99(2), 223–233.

<https://doi.org/10.1111/j.1365-2672.2005.02596.x>

Moller, A. G., Winston, K., Ji, S., Wang, J., Davis, M. N., Solis-Lemus, C. R., & Read, T.

(2020). Genes Influencing Phage Host Range in *Staphylococcus aureus* on a Species-Wide Scale. BioRxiv. doi:10.1101/2020.07.24.218685

Norian, H. A. A Bioinformatics Approach to Revealing the Genetic Basis for Host Range

Specificity (2014-2015). James Madison Undergraduate Research Journal, 2(1).

<https://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1008&context=jmurj>

N Saitou, M Nei, The neighbor-joining method: a new method for reconstructing

phylogenetic trees., Molecular Biology and Evolution, Volume 4, Issue 4, Jul

1987, Pages 406–425, <https://doi.org/10.1093/oxfordjournals.molbev.a040454>

Oscar Robinson, David Dylus, Christophe Dessimoz; Phylo.io : Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web . Mol Biol Evol 2016; 33 (8): 2163-2166. doi: 10.1093/molbev/msw080

Pope WH, Ferreira CM, Jacobs-Sera D, Benjamin RC, Davis AJ, DeJong RJ, Elgin SC, Guilfoile FR, Forsyth MH, Harris AD, Harvey SE, Hughes LE, Hynes PM, Jackson AS, Jalal MD, MacMurray EA, Manley CM, McDonough MJ, Mosier JL, Osterbann LJ, Rabinowitz HS, Rhyan CN, Russell DA, Saha MS, Shaffer CD, Simon SE, Sims EF, Tovar IG, Weisser EG, Wertz JT, Weston-Hafer KA, Williamson KE, Zhang B, Cresawn SG, Jain P, Piuri M, Jacobs WR Jr, Hendrix RW, Hatfull GF. Cluster K Mycobacteriophages: Insights into the Evolutionary Origins of Mycobacteriophage TM4. PLoS One. 2011;6(10):e26750. doi: 10.1371/journal.pone.0026750. Epub 2011 Oct 28. PMID: 22053209; PMCID: PMC3203893.

Potera C. 2013. Phage Renaissance: New Hope Against Antibiotic Resistance. Environ Health Perspect 121:a48–a53. doi:10.1289/ehp.121-a48.

Ross, A., Ward, S., & Hyman, P. (2016). More Is Better: Selecting for Broad Host Range Bacteriophages. Frontiers in Microbiology, 7, 1352.
<https://doi.org/10.3389/fmicb.2016.01352>

Russell, D. A., & Hatfull, G. F. (2017). PhagesDB: the Actinobacteriophage Database. Bioinformatics (Oxford, England), 33(5), 784–786.
<https://doi.org/10.1093/bioinformatics/btw711>

Smith KC, Castro-Nallar E, Fisher JN, Breakwell DP, Grose JH, Burnett SH: Phage Cluster Relationships Identified through Single Gene Analysis. BMC Genomics 2013, 14:410.

Sudhir Kumar, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura (2018)

MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms.

Molecular Biology and Evolution 35:1547-1549

Twort, F. W. (1915). An Investigation on the Nature of Ultra-microscopic Viruses. The

Lancet, 186(4814), 1241-1243.