Conserved Repetitive Sequences within Mycobacteriophage Giles Suggest Illegitimate Recombination Events between Phage and Bacteria

**Introduction**

 Bacteriophages, viruses which infect bacteria, are the most abundant organic agent on the planet, sporting a world population of an estimated 1031 phages (Morris, 2008). Given such a large population, phage genomes are highly diverse and mosaic in nature. Genes within one phage can be found to be shared among other phages, and host-phage interactions can result in illegitimate recombination events. Some instances of host-phage illegitimate recombination can effect a bacteria’s resistance against phage infection and proliferation, as in the production of Cas proteins (Haft, 2005).

The genome of mycobacteriophage Giles is a particularly interesting case of genetic diversity. The phylogenic relationships between phages are sorted and documented within groups called clusters (Hatful, 2010). Giles is a phage belonging to a particularly unique cluster called cluster Q, which consists of five known bacteriophages, all of which are nearly genetically identical (Cresawn et al., 2011). Indeed, the genetic database software Phamerator records for these phages indicate less than 28% of genes in this cluster are related to other known bacteriophages, making the phages within cluster Q not only genetically identical, but also genetically unique. Genes in Giles conserved in other phages exist in an abundance of other clusters, making Giles a very diverse phage in terms of possible phylogenic origins of its highly mosaic genome.

 Adjacent repetitive sequences within DNA, known as tandem repeats, are a phenomena well-studied within the realm of sequence differentiation and diversity in mycobacteria, but little is known about their presence within the phages that infect them (Sun, 2004). Giles’ genome was studied for the presence of possible tandem repeats within its genes, and repeat sequences within this phage were identified. Other mycobacteriophages were searched for the presence of like sequences to determine possible recombination events, and the genetic content of mycobacteria were likewise compared to Giles’ repeat sequences to identify possible instances of illegitimate recombination between phage and host.

**Methods**

Tandem repeats in mycobacteriophage Giles were identified through an autocorrelation function programmed in BioBike, using the Phantome phage and bacteria database (Elhai, 2009). This function parses through each nucleotide coordinate in the phage genome, and compares each seventh nucleotide to the seventh nucleotide preceding. For instance, in the repeat region of CGTCTGACGTCTGA, the first and eight nucleotides are compared for identity match, and the identical nucleotide of “C” at both coordinates would mean the function count is incremented by one. The second and ninth nucleotides are then likewise compared, and the function count is once more incremented based on matching nucleotide content. This process is repeated within a window of 100 nucleotides, and the window region is then assigned a score ranging from one to 100; a score of one would indicate a total of one nucleotide match within the 100 nucleotide window, while a score of 100 indicates a string of entirely identical nucleotides. The probability of a repeated unit over 40 within the perimeter of four possible nucleotides and a 100 nucleotide window is calculated as to occur randomly only 0.04% of the time, making this an appropriate cutoff threshold for identity score matches per window.

**Figure 1. Illustration of BioBIKE autocorrelation window.** Each seventh nucleotide in this 14 nucleotide sequence is compared for identity match. The first sequence contains a dual repeat of a single sequence, giving it a total score of seven. The second sequence contains a mismatch between the fourth nucleotide and its corresponding unit seven nucleotides ahead, making its total match score six.

 Once probable regions of tandem repeats were found, the genes containing these repeats were documented, and phage database software Phamerator was used to identify other mycobacteriophages that conserve these genes within their cluster. This data was cross-examined against a “sequence/s-similar-to” function of all mycobacteriophage genomes on BioBIKE in order to confirm the presence or absence of these sequences in other phages. A “sequence/s-similar-to” function was also used to identify similar sequences within all mycobacteriophages and mycobacteria in the database.

**Results**

In the entirety of Giles’ DNA sequence, two regions were found to have a number of matching repeated units over 40 (Figure 3). The coordinates corresponding to these peaks were identified as being located in Giles gene products 27 and 46. Further analysis of these regions determined the exact sequence of the repeats, information which can be found in Figures 4 and 5.

 The “sequence/s-similar-to” function resulted in no matches of these repeat sequences within other mycobacteriophages. However, these sequences were found to occur with mycobacteria. Match information is available in Figures 4 and 5.

**Figure 3. Coordinates with number of repeated units over 40.** These coordinates were found to correlate with gp27 and gp46 in Giles.



**Figure 4. Repeat sequence in gp46 and match in mycobacteria.** The percent identity match in *M. jls* is of a notably lower quality than in *M. Smegmatis*, and its higher e-value score may indicate this match was simply a random occurrence.





**Figure 5. Repeat sequence in gp46 and match in mycobacteria.**

**Discussion.**

The repetitive sequences in phage Giles’ genes 27 and 46 were not found as conserved in any other mycobacteriophage. Indeed, Phamerator data indicates the entirety of these genes are wholly unique to cluster Q, and has no identified function for either of these proteins (Cresawn et al., 2011). According to protein interaction data compiled by Mehla *et al.*, Giles gene 27 is a gene product associated with tail assembly (Mehla *et al.,* 2015). Gp46 has no known documented function within the phage, and has only one apparent laboratory protein interaction with minor tail protein gp17.

Interestingly, though no phage conservation of this sequence was found, the repeat sequences in both gp27 and gp46 were found to exist within mycobacteria. Full host range data for Giles is not completely known, but the phage is able to infect *M. smegmatis*, and is not able to form plaques when plated with *M. tuberculosis* (Dedrick et al., 2013). The similar sequence match of a sequence in Giles gp27 within *M. tuberculosis* strains F11 and CDC1551 may suggest an event of illegitimate recombination, and the existence of this sequence may account for *M. tuberculosis’s* apparent immunity to the phage. The same sequence was likewise found in two strains of *M. bovis*, and further host range studies may suggest this sequence’s role in the prevention of phage proliferation within certain mycobacteria.

**Figure 6. Phamerator genome map for mycobacteriophage Giles.** Rectangles represent identified genes within the phage, boxed letters indicate clusters where genetically similar genes are found. Circled is identified gene 46, a gene of unidentified function and exclusively conserved within cluster Q.

An illegitimate recombination event appears to be present in gp46. The same sequence is found in *M. smegmatis* mc2, and a lesser quality match in *M. jls*. Though host range data on Giles is limited in capacity, the lack of significant interaction data and known nonessentiality of that protein as determined by Mehla *et al.* suggests it may have originated within the host bacteria and crossed over into Giles’ genome. Further host range studies may confirm this result.

**References**

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.

Dedrick RM, Marinelli LJ, Newton GL, Pogliano K, Pogliano J, Hatfull GF. Functional requirements for bacteriophage growth: gene essentiality and expression in mycobacteriophage Giles. Mol Microbiol. 2013;88(3):577-89.

Elhai J, Taton A, Massar JP, et al. BioBIKE: a Web-based, programmable, integrated biological knowledge base. Nucleic Acids Res. 2009;37(Web Server issue):W28-32.

Haft DH, Selengut J, Mongodin EF, Nelson KE. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. 2005;1(6):e60.

Hatfull GF, Jacobs-sera D, Lawrence JG, et al. Comparative genomic analysis of 60 Mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. J Mol Biol. 2010;397(1):119-43.

\*Manuscript headed by Jitender Mehla currently in revision process for Journal of Bacteriology (initial reviews were positive to the results of the paper – not certain how to cite this given it is not published in a public database yet.)

Morris P, Marinelli LJ, Jacobs-sera D, Hendrix RW, Hatfull GF. Genomic characterization of mycobacteriophage Giles: evidence for phage acquisition of host DNA by illegitimate recombination. J Bacteriol. 2008;190(6):2172-82.

Sun YJ, Bellamy R, Lee AS, et al. Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of Mycobacterium tuberculosis in Singapore. J Clin Microbiol. 2004;42(5):1986-93.