INSTITUT PASTEUR

Research in Microbiology 159 (2008) 332–339

www.elsevier.com/locate/resmic

# Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution

Graham F. Hatfull [a,*], Steven G. Cresawn [b], Roger W. Hendrix [a]

[a] *Department of Biological Sciences and Pittsburgh Bacteriophage Institute, University of Pittsburgh, Pittsburgh, PA 15260, USA*
[b] *Department of Biology, James Madison University, Harrisonburg, VA 22807, USA*

Received 12 March 2008; accepted 14 April 2008
Available online 7 May 2008

**Abstract**

The recognition of the vast numbers of bacteriophages in the biosphere has prompted a renewal of interest in understanding their morphological and genetic diversity, and elucidating the evolutionary mechanisms that give rise to them. We have approached these questions by isolating and characterizing a collection of mycobacteriophages that infect a common bacterial host, *Mycobacterium smegmatis*. Comparative genomic analysis of 50 mycobacteriophages shows that they are highly diverse, although not uniformly so, that they are pervasively mosaic with a multitude of single gene modules, and that this mosaicism is generated through illegitimate recombination.
© 2008 Elsevier Masson SAS. All rights reserved.

*Keywords:* *Mycobacterium*; Virion morphotype; Siphoviridae; Myoviridae; Genome mosaicism

## 1. Introduction

Mycobacteriophages are viruses that infect mycobacterial hosts. While there are more than 60 different species in the genus *Mycobacterium*, two stand out as prominent human pathogens: *Mycobacterium tuberculosis*, the cause of human tuberculosis, and *Mycobacterium leprae*, the cause of leprosy (for review, see [22]). In spite of the medical importance of these pathogens, there was little progress made in their genetic analysis until the late 1980s, primarily because of the lack of genetic tools and approaches [23]. However, an additional problem with *M. leprae* is that it never has convincingly been shown to grow in defined in vitro culture conditions, although it can be propagated in armadillos and mouse footpads [29,36]. In contrast, *M. tuberculosis* can be readily grown in defined media in vitro, although its extremely slow growth rate (24 h doubling time) and its pathogenicity present serious challenges to the microbial geneticist [23].

Advancement of mycobacterial genetics has been enormously stimulated by isolation and characterization of mycobacteriophages [16]. The first mycobacteriophages were isolated more than 60 years ago, and were developed soon after as phage-typing tools for diagnostic purposes [35]. More recently, we and others have expanded the collection of mycobacteriophages by isolating new phages from the environment using *Mycobacterium smegmatis* as a host, and characterizing them by genomic analysis [17]. This effort serves two key purposes. The first is to provide an array of genetic tools for dissection and understanding of mycobacterial hosts [15]. The second is to provide insights into viral diversity and evolution through the comparative genomics of a group of viruses that all infect a common host [17]. Here we focus on the latter and review what lessons we have learned from the comparative genomic analysis of more than 50 mycobacteriophages with regard to their genetic diversity and the evolutionary processes that give rise to them. In general, we would predict that these lessons are applicable to other groups of bacteriophages that infect a common host, and that there will be few systematic idiosyncrasies of the mycobacteriophages that make them different from other phages. One exception to this concerns the viral–host interactions and DNA injection, since the

* Corresponding author.
*E-mail address:* gfh@pitt.edu (G.F. Hatfull).

0923-2508/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.resmic.2008.04.008

mycobacteria possess complex lipid-rich cell walls that are relatively uncommon among bacterial hosts [3].

## 2. The mycobacteriophage collection

Several hundred mycobacteriophages have been reported in the literature, many isolated for phage typing purposes. Relatively few have been characterized in detail, including genomic characterization; mycobacteriophages L5 and D29 were isolated in Japan and in the US respectively [10,12], and TM4 was recovered from a putative lysogenic strain of *Mycobacterium avium* [39]. All of the other sequenced genomes are from more recently isolated phages and were recovered from environmental samples (e.g. soil, compost, etc.) using *M. smegmatis* as a host and without any step for viral amplification [17,32]. A total of more than 50 genomes have been sequenced, although we focus primarily here on the comparative analysis of 32 genomes that we have examined most closely [17,31,33]. Little is known about the host range of these phages, although approximately 10% can infect *M. tuberculosis*. All of these mycobacteriophages contain large (>40 kbp) dsDNA genomes (Table 1).

Table 1
Features of completely sequenced mycobacteriophages

| Phage | Morphotype | GC% | Size (kbp) | orfs | tRNAs | tmRNAs |
|---|---|---|---|---|---|---|
| 244 | Siphoviridae | 62.9 | 74.5 | 142 | 2 | 0 |
| Barnyard | Siphoviridae | 57.3 | 70.8 | 109 | 0 | 0 |
| Bethlehem | Siphoviridae | 63.2 | 52.2 | 87 | 0 | 0 |
| Bxb1 | Siphoviridae | 63.6 | 50.5 | 86 | 0 | 0 |
| Bxz1 | Myoviridae | 64.8 | 156.1 | 225 | 30 | 1 |
| Bxz2 | Siphoviridae | 64.2 | 50.9 | 86 | 3 | 0 |
| Catera | Myoviridae | 64.7 | 153.8 | 218 | 29 | 1 |
| Che8 | Siphoviridae | 61.3 | 49.5 | 112 | 0 | 0 |
| Che9c | Siphoviridae | 65.4 | 57.5 | 84 | 0 | 0 |
| Che9d | Siphoviridae | 60.9 | 56.2 | 111 | 0 | 0 |
| Che12 | Siphoviridae | 62.9 | 52.0 | 98 | 3 | 0 |
| CJW1 | Siphoviridae | 63.1 | 76.0 | 141 | 2 | 0 |
| Cooper | Siphoviridae | 69.1 | 70.6 | 99 | 0 | 0 |
| Corndog | Siphoviridae | 65.4 | 69.8 | 122 | 0 | 0 |
| D29 | Siphoviridae | 63.5 | 49.1 | 77 | 5 | 0 |
| Giles | Siphoviridae | 67.3 | 54.5 | 79 | 0 | 0 |
| Halo | Siphoviridae | 66.7 | 42.3 | 65 | 0 | 0 |
| L5 | Siphoviridae | 62.3 | 52.3 | 90 | 3 | 0 |
| LLIJ | Siphoviridae | 61.5 | 56.8 | 100 | 0 | 0 |
| Omega | Siphoviridae | 61.4 | 110.9 | 237 | 2 | 0 |
| Orion | Siphoviridae | 66.5 | 68.4 | 100 | 0 | 0 |
| PBI1 | Siphoviridae | 59.7 | 64.5 | 81 | 0 | 0 |
| PG1 | Siphoviridae | 66.5 | 69.0 | 100 | 0 | 0 |
| Pipefish | Siphoviridae | 67.3 | 69.0 | 102 | 0 | 0 |
| P-Lot | Siphoviridae | 59.7 | 64.8 | 89 | 0 | 0 |
| PMC | Siphoviridae | 61.4 | 56.7 | 104 | 0 | 0 |
| Qyrzula | Siphoviridae | 69.0 | 67.2 | 81 | 0 | 0 |
| Rosebush | Siphoviridae | 69.0 | 67.5 | 90 | 0 | 0 |
| TM4 | Siphoviridae | 68.1 | 52.8 | 92 | 0 | 0 |
| Tweety | Siphoviridae | 61.7 | 58.7 | 109 | 0 | 0 |
| U2 | Siphoviridae | 63.7 | 51.3 | 81 | 0 | 0 |
| Wildcat | Siphoviridae | 56.9 | 78.3 | 148 | 22 | 1 |
| Average | | 63.8 | 68.0 | 111 | | |
| Total | | | 2175 | 3545 | 101 | 3 |

## 3. Mycobacteriophage viral morphologies

To our knowledge, all of the mycobacteriophages isolated to date are tailed dsDNA phages, and this is certainly true for all of those whose genomes have been sequenced. Three morphotypes of dsDNA tailed phages have been described previously according to the nature of the phage tail – the myoviridae that contain contractile tails, the podoviridae with short stubby tails, and the siphoviridae with long flexible non-contractile tails [1]. Interestingly, we are not aware of a single mycobacteriophage of the podoviral morphotype and all appear to be either of the myoviridae or siphoviridae types; of 32 published sequenced genomes, only two are myoviral and all of the others are siphoviral (Fig. 1). The reasons for the absence of podoviral phage in this collection is not clear, although as more mycobacteriophages are characterized, it seems unlikely that this is due to sampling variations. We speculate that phages with extremely short tails are unable to navigate the mycobacterial cell wall and associate with the cell membrane. Most of the mycobacteriophages have isometric heads, although some (e.g. Che9c, Corndog) have extensively elongated heads (Fig. 1).

## 4. Mycobacteriophage genometrics

The genome length of mycobacteriophages varies substantially, with the shortest being ~42 kbp and the largest ~150 kbp (Table 1). However, the average length is ~70 kbp and the sizes are not restricted to a few well-defined groups. The two largest phages are Bxz1 and Catera and both have a myoviral morphotype. The lengths of the genomes of siphoviral phages vary from 42 kbp to 110 kbp with many different sizes within this range [17,32]. We note that the ~70 kbp average genome length is about twice that of the sequenced dairy phages [5] and it is unclear what the determinants of genome length are, although at least for genomes smaller than a 100 kbp, there is some correlation between GC% and length [32].

The average GC% content of the mycobacteriophages (63.7%) is similar to that of the host *M. smegmatis* (63%), although there is also substantial variation in GC%, ranging from 59% (mycobacteriophage Wildcat) to 69% (e.g. mycobacteriophage Rosebush) (Table 1). One possible explanation to account for this variation is that these phages may have substantially different host ranges, and that the GC% reflects that of their preferred bacterial hosts in their natural environment, even though they can all infect the *M. smegmatis* host that was used to isolate them. As noted above, comprehensive host range analysis has not yet been explored.

There is also substantial variation in the tRNA gene content of the mycobacteriophages (Table 1). About two-thirds of the genomes have no tRNA genes, whereas others have as many as 30 or more, although those with a larger number are typically those with a bigger genome and a myoviral morphotype (Table 1). There are many genomes with just a handful of tRNA genes, and a variety of anticodon and tRNA types are represented. The tRNA content can also vary between very

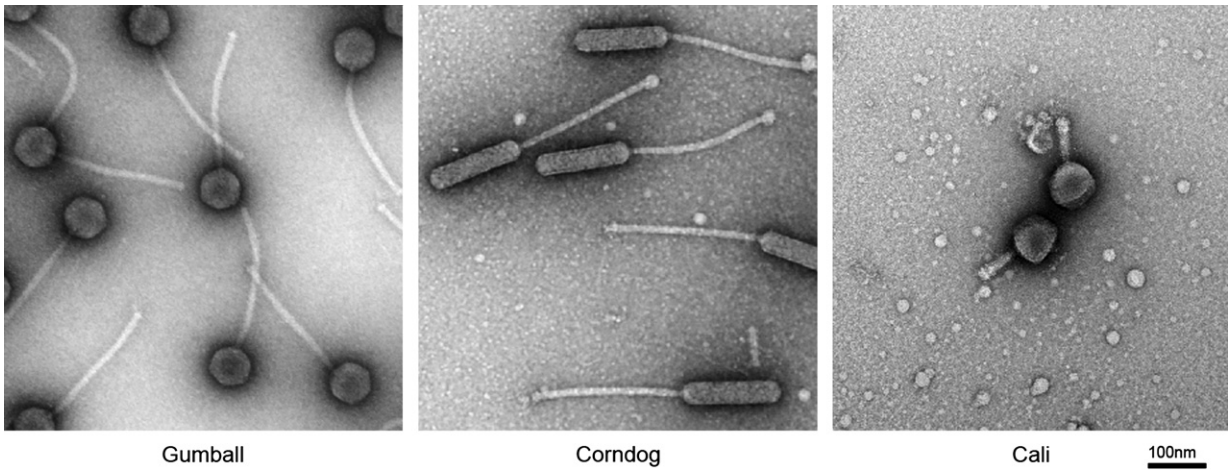


Fig. 1. Mycobacteriophage virion morphotypes. Mycobacteriophages generally belong to either the siphoviridae (e.g. Gumball, Corndog) or myoviridae (e.g. Cali) morphotypes, and no examples of the podoviridae have been described. Those within the siphoviridae group have either an isometric head as in Gumball, or an elongated head as in Corndog.

closely related phages such as L5 and D29 which share substantial nucleotide sequence similarity (~80% nucleotide identity) in their left arms; D29 has a cluster of five tRNA genes while L5 has only three in the same genomic location [11,18]. The pattern of tRNA gene distribution is thus not particularly helpful in elucidating their roles in bacteriophage evolution, and the reason why they are acquired remains unclear [34].

## 5. Clustering of mycobacteriophage genomes by nucleotide similarity

Nucleotide sequence comparison of 30 mycobacteriophage genomes reveals two main features (Fig. 2). First, there is evidently a high degree of genetic diversity; these are not simply all minor variations of one or two common sequences [17]. The diversity is amazing given that all of these phages infect at least one common host (i.e. *M. smegmatis*) and are therefore expected to be in direct genetic communication with each other. Overall, the diversity appears to be generally greater than among the phage collections of either *Pseudomonas* or *Staphylococcus* [27,28]. However, it is not clear if this is really reflective of a difference among the population of phages that infect these different hosts, if it reflects the complexity of the diversity of different host bacterial populations, or if it simply results from differences in the phage isolation procedures.

The second feature is that mycobacteriophage genetic diversity is not uniform, and there are clusters of genomes that appear to be more closely related to phages within the cluster than to those outside of it (Fig. 2). Six specific clusters can be identified, although it should be noted that genomes within a cluster can be very similar (e.g. Bxz1 and Catera share >90% nucleotide sequence identity) or can be more distantly related, with either weak overall sequence similarity, or having regions of evident sequence commonality interrupted by genome segments that are much more distantly related. Inclusion of genomes within clusters is thus largely on an ad hoc basis rather than according to strict criteria. For example, while there is little doubt that Cjw1 and 244 can be clustered together, it is not clear whether Omega should be included within the cluster containing PMC, Che8 etc., since the similarity is primarily restricted to just subparts of the genomes (Fig. 2). Thus clustering is unlikely to reflect true subdivisions within the population, and is expected to be modified significantly as more genome sequences become available. In this regard, given the overall diversity of the mycobacteriophages, the total number of sequenced genomes available is too small to provide much insight into the overall structure of the larger population of these phages.

## 6. Mycobacteriophage genetic mosaicism

The complex genetic relationships among the mycobacteriophages reflected at the nucleotide sequence level show that some segments of the genomes must have different evolutionary histories from others [21]. This concept of phage genomes being composed of different segments with non-congruent phylogenetic relationships is supported by comparative genomics of all known phages and reflects the strong contribution of horizontal genetic exchange events [20]. However, comparison at the nucleotide level is only of limited use in comparing the mycobacteriophages because of their substantial diversity. Further insights can be gained by comparing the amino acid sequences of the predicted proteins, where more distant signs of common ancestry are present, even when any sequence commonality is lost at the nucleotide level.

To facilitate this analysis we have developed a program 'Phamerator' that assorts the ~3300 predicted proteins encoded by 30 mycobacteriophages into 'phamilies' of sequences that are related by a Blast score of 0.0001 or better, or with greater than 27.5% amino acid sequence identity (our unpublished data), and simplifies a previously described non-computational analysis [17]. The analysis is complicated by the presence of a number of genes that are chimeric, and

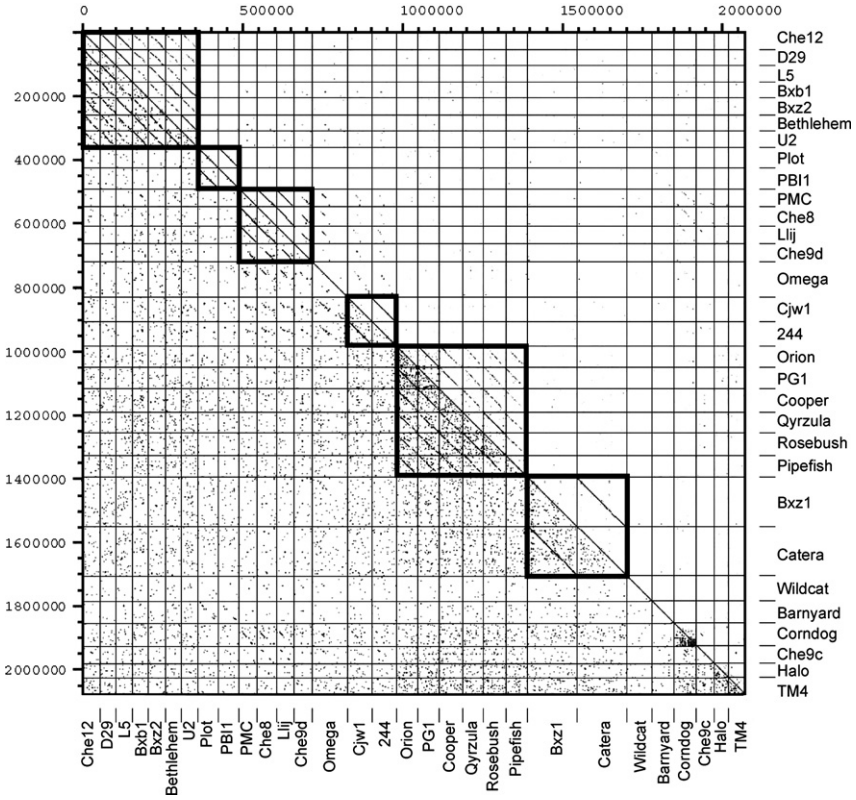


Fig. 2. Clustering of mycobacteriophage genomes by nucleotide sequence similarity. Dotter analysis comparing 30 mycobacteriophage genomes is displayed with boxes showing those genomes that appear to be more closely related to genomes within the cluster than to those outside of it. The Dotter output in the lower left triangle reports a more relaxed sequence comparison and the upper right is more stringent.

composed of gene segments with different evolutionary histories. This leads to the generation of a few 'superphamilies' composed of large numbers of genes, but in which not all pairwise comparisons exhibit similarities. Perhaps the best example is a large phamily of structural proteins which contains more than 300 genes! Most of these correspond to minor tail proteins, and efforts to de-convolute the superphamily into conserved subdomains are complex because the boundaries are ill-defined and cannot be assigned to a relatively small number of positions where recombination occurs [17]. We assume that this reflects the structural properties of these tail proteins that predominantly form extended rather than globular structures.

The grouping of mycobacteriophage genes into phamilies is instructive in a multitude of ways. First, the number of phamilies (approx. 1400) shows that a very large number of different sequences are represented, reflecting the genetic diversity seen by nucleotide sequence comparison (Fig. 2). Second, more than 50% of these phamilies contain only a single gene member, again supporting the high genetic diversity, but suggesting that there is likely to also be an abundance of novel genes that have not been previously described [17]. Third, only a small proportion (~15%) of all of the phamilies have relatives in genomes other than mycobacteriophages (i.e. in the sequence databases). Taken together with the estimate that there are $10^{31}$ phage particles globally, these observations confirm that phages in general are likely to contain the largest reservoir of unexplored sequences in all of nature [17]. Only about one-half of those phamilies that match database entries are related to other phage genes, with the other half related to bacterial genes [17].

The assortment of genes into phamilies is especially helpful for representing genome mosaicism. For example, the Phamerator program can simply generate genome maps in which each gene is color-coded and named for the pham that it belongs to; a small segment corresponding to genes *86–90* of Che9d is shown in Fig. 3. Furthermore, Phamerator can also generate 'phamily circles' which offer an alternative representation of the evolutionary histories of these genes (Fig. 3). The advantage of these phamily circle representations is that they enable a simple means of comparing different phamilies within a genomic context. This is difficult using traditional phylogenetic trees because in most cases the homologues are not found in the same set of genomes [17]. In phamily circles, all of the genomes being compared are represented, with arcs noting the presence of related genes with the thickness of the line corresponding to the strength of the similarity. In Fig. 3 for example, it can be clearly seen that the evolutionary histories of the five consecutive genes (Che9c *86–90*) are all distinctly different to each other. The finding that the mosaic modules often correspond to just single genes is a common feature of the mycobacteriophages, especially outside of the virion structure and assembly operons.

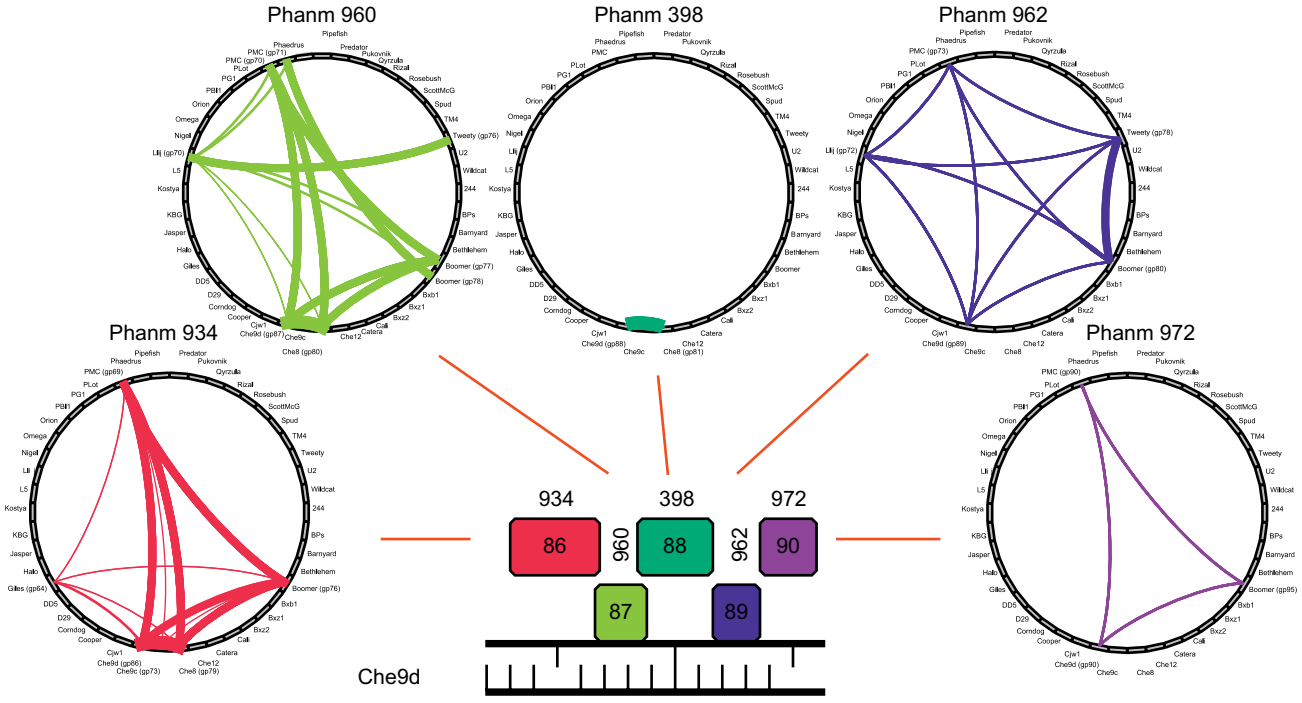


Fig. 3. Illustration of mycobacteriophage genomic mosaicism. The center of the figure shows a small segment of the Che9d genome containing genes *86–90*. The genes are color-coded according to the phamilies to which they belong and the phamily numbers are shown above each gene. The phamily circles for each of these genes show how other phamily gene members distribute among other mycobacteriophage genomes. It is clear from this representation that these five consecutive Che9d genes all have distinct phylogenetic histories and thus correspond to separate modules constituting an overall mosaic structure.

## 7. The role of illegitimate recombination in generating genome mosaicism

It was noted in early comparisons of bacteriophage genomes that mosaic boundaries appear to correspond closely to borders between genes [6,38]. There are examples of this also in mycobacteriophages where phages may show close similarity at the DNA sequence level but the similarity has been interrupted through a horizontal genetic exchange event. One such example is revealed by the comparisons of phages Cjw1 and 244. Overall, these genomes are very similar and cluster together in the nucleotide sequence analysis (Fig. 2). However, an alignment of the maps shows that there is pham synteny for most of the genes in the *79–91* intervals, with the exception of 244 *86* which is unrelated to Cjw1 *85* (Fig. 4A). The site of sequence departure at both boundaries is precisely at the gene boundaries, and the right boundary is shown in Fig. 4B). This comparison highlights an additional aspect of comparative genomic approaches, in that the alignment suggests strongly that the unusual UUG codon is for translation initiation of 244 gene *87* (Fig. 4B), corroborating previous reports of UUG as a start codon in mycobacteriophage L5 [18].

There are two alternative models to account for the molecular events involved in horizontal genetic exchange of phage genomes and the creation of mosaicism. An early model by Susskind and Botstein [38] proposed that short conserved sequences at gene boundaries (i.e. boundary sequences) could serve to target the homologous recombination machinery to gene borders. However, while there are examples supporting this model in some genomes [9], it is not obviously involved in generating mycobacteriophage mosaicism, and conserved boundary sequences are not common. An alternative model proposes that mosaicism is generated primarily by illegitimate recombination without extensive sequence similarity, and essentially at random [7,20,25]. Most of these events would give rise to non-viable phage genomes that are of the wrong length or have lost essential functions. In rare circumstances, the events (perhaps involving multiple recombination events) may lead to viable progeny and propagation of one or more new recombinant boundaries. The illegitimate recombination events at gene boundaries are more likely to be evolutionarily productive by retaining gene functions.

Unfortunately, a genome comparison such as that shown in Fig. 4B does not readily help to distinguish between these two models, and is consistent with both. However, a second type of rearrangement emerges from the comparative genomics of mycobacteriophages, and one example is present in the genome of mycobacteriophage Giles [31]. At the right end of the Giles genome, the rightmost gene is *79*, a 303 bp open reading frame whose predicted amino acid sequence is strongly related to the *M. smegmatis* MetE protein. Comparison at the nucleotide sequence level shows that there is close similarity between a 203 bp region of the Giles genome and bacterial *metE* genes, with 100% identity with *metE* of *M. smegmatis*. This corresponds to only a small part of the *metE* gene which is over 2 kbp long, and it is unclear whether Giles *79* corresponds to a functional domain or is simply non-functional. The high level of sequence similarity suggests that this DNA was acquired from either *M. smegmatis* itself or

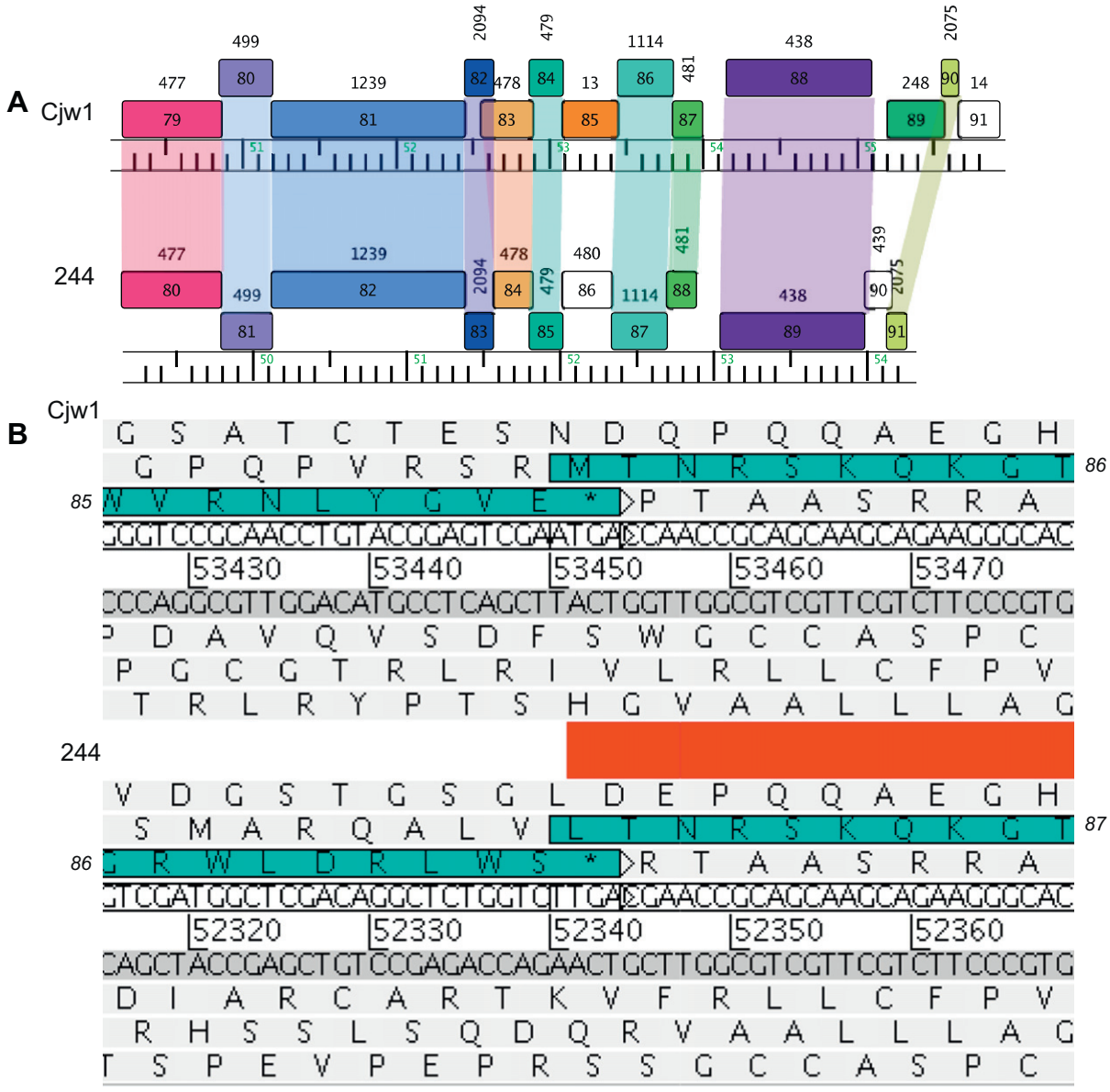


Fig. 4. Mycobacteriophage genome mosaicism and gene boundaries. (A) Alignment of Cjw1 genes *79–91* and 244 genes *80–91*. These genomes cluster together by similarity of their nucleotide sequences, and the genes in this region are generally syntenic and near-identical in sequence. Cjw1 gene *85* and 244 gene *86* represent a syntenic break and are completely unrelated genes. (B) Alignment of the Cjw1 and 244 sequences at the boundaries of Cjw1 genes *85–86* and 244 genes *86–87* shows that the common sequences (shown in red) extend just to the second base of the start codon of the related genes. The left junction (between Cjw1 genes 84–85 and 244 gene *85–86*) also occurs at the gene boundaries (data not shown).

from a very closely related but as yet uncharacterized host, and that this acquisition occurred relatively recently [31]. However, there is no simple explanation to account for its acquisition by any targeted recombination mechanism and it seems probable that it was acquired by an illegitimate recombination event.

## 8. Genome architectures and translocations

Notwithstanding the high degree of mycobacteriophage genome diversity, among those phages with a long flexible non-contractile tail there are common features of their genome architectures. The most prominent of these is the common arrangement of the virion structure and assembly genes [8,17,32]. The gene order of the >20 genes is well conserved in most bacteriophages of this type, even though the sequences may be quite different, and additional genes (morons [25]) may interrupt what are otherwise strongly syntenic relationships. In a rather extreme case, in phage Omega this virion structural gene operon spans ~36 kbp, a substantially larger segment than the equivalent operon (~21 kbp) in phage Halo [32].

All of the sequenced mycobacteriophage genomes encode a lysis cassette that typically contains two putative lytic enzyme genes (*lysA* and *lysB*) [13]; in a subset of genomes a putative holin gene can also be recognized. However, this cassette is found in several different locations relative to the virion structural gene operon. In some phages (e.g. L5 and

closely related phages) the lysis cassette is located immediately to the left of the terminase gene and lies in between terminase and the putative *cos* packaging site [18]. But in many other genomes (e.g. Tweety [33]) the lysis cassette is to the right of the virion structural genes, typically between the tail protein genes and the integration functions. There are not yet any genome comparisons that reveal a relatively recent switch of this cassette from one position to the other.

The integration functions usually include an integrase gene, an *attP* attachment site and a recombination directionality factor (RDF, or Xis) gene. Both tyrosine-integrase (lambda-like) and serine-integrase (resolvase-like) genes are represented and since the tyrosine-integrases typically utilize a host *attB* integration site that overlaps a tRNA gene, the phage *attP* site often can also be identified [33]. The attachment sites for the serine-integrases are small and typically must be determined experimentally [26]. RDF genes are highly diverse and can be recognized in some genomes but not in others [14,30]. The larger collection of about 50 mycobacteriophage genomes shows that about half have one of these integration systems.

The integration cassettes are typically positioned close to the midpoint of the genomes of those with defined cohesive termini, to the right of the virion structural gene operon. There is one notable exception, which is in mycobacteriophage Giles [31]. Here, the integration cassette appears to have migrated from its more typical position to a location within the structural gene operon [31]. This is evidenced by the observation that the lysis cassette and a structural gene lie to the right of the integration cassette while other head and tail genes are to the left [31]. The simple interpretation is that the integration functions have translocated to the new position via an integrase-mediated recombination event using secondary or non-canonical attachment sites [31]. Thus genome translocations likely occur both by illegitimate and site-specific recombination processes.

## 9. Mycobacteriophage recombination functions

The proposal that illegitimate recombination plays a key role in generating genomic mosaicism assumes that these events are infrequent but are evolutionarily creative. Because there is a high frequency of phage infections (estimated at $10^{25}$ infections/second globally) and phage evolution has likely been proceeding for more than three billion years [19], there is no need to suppose that specific phage-encoded recombinases are involved. Presumably any recombinase that promotes homologous recombination can also do so non-homologously albeit at a much lower frequency. It is of interest to note that specific processes have been described that could contribute to illegitimate recombination, with the most compelling being the ill-defined mechanism that gives rise to the acquisition of CRISPR sequences implicated in conferring resistance to phage infection [2,37]. The direct repeat locus of *M. tuberculosis* has similarities to CRISPR sequences [4], but it is not yet clear if these are related to any of the mycobacteriophage sequences.

Genes encoding homologous recombination functions have been identified in several mycobacteriophage genomes, although these are of several distinct classes. First, nine mycobacteriophages encode a *recA* homologue, and these include those phages in the Bxz1 and Cjw1 clusters. All of these genomes assemble as circular molecules and we presume that the genome ends are terminally redundant and circularly permuted. One potential role of the phage-encoded RecA proteins is that they mediate genome circularization following infection, particularly if host RecA levels are low, and we note that all of these phages are capable of infecting *recA*$^-$ mutants of *M. smegmatis* (T. Sampson and G.F.H., unpublished observations). We do not yet know whether *recA* is essential for propagation of these phages.

Several mycobacteriophages encode proteins related to the *recE/T* system of *E. coli*. Che9c *60* and *61* are clearly homologues of *E. coli recE* and *recT* respectively and have been determined biochemically to perform similar functions [40]; mycobacteriophages Halo and Giles appear to have variations of these genes. Che9c gp60 and gp61 confer elevated recombination frequencies with linear dsDNA substrates, and gp61 alone can be used to recombine *M. smegmatis* and *M. tuberculosis* with ssDNA oligonucleotide substrates [41]. Che9c gp61-mediated recombination is very efficient and requires only a relatively small (>40 nuc) ssDNA substrate [41]. Other studies suggest that other mycobacteriophages (e.g. TM4) may also have recombination systems, even though no recombination genes have been identified in their genomes [24]. We predict that many mycobacteriophages encode as yet unrecognized recombination systems and note that many (e.g. TM4) do encode identifiable Holliday junction resolving enzymes.

## 10. Concluding remarks

In summary, the mycobacteriophages are a remarkably diverse group of viruses whose characterization has provided helpful insights into the mosaic nature of bacteriophage genomes and the evolutionary mechanisms that give rise to them. The observed diversity is especially intriguing given that all these phages infect a common host, although we speculate that they have different but overlapping host ranges which are reflected in their different GC% contents. Nevertheless, with more than 50 complete mycobacteriophage genome sequences available it is clear that we are far from saturating the known mycobacteriophage types or the known mycobacteriophage genes. Thus, while it becomes increasingly likely that a newly characterized mycobacteriophage will have some nucleotide similarity to a phage in the current collection, phages such as Giles that have no close relatives and in which more than 50% of the genes are novel will no doubt still be isolated. Further expansion of our knowledge of mycobacteriophage genomics is thus expected to continue to advance our understanding of their diversity and evolution. But it is becoming increasingly clear that our insights into mycobacteriophage genomes are constrained by our lack of understanding of the functions of more than 1000 new sequence phamilies, and we hope that functional genomic approaches will help

to provide answers to some of the questions regarding what these mycobacteriophage genes do.

## References


[1] Ackermann, H.W. (2007) 5500 Phages examined in the electron microscope. Arch. Virol. 152, 227–243.

[2] Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., Horvath, P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315, 1709–1712.

[3] Brennan, P.J. (2003) Structure, function, and biogenesis of the cell wall of *Mycobacterium tuberculosis*. Tuberculosis (Edinb.) 83, 91–97.

[4] Brudey, K., Driscoll, J.R., Rigouts, L., Prodinger, W.M., Gori, A., Al-Hajoj, S.A., Allix, C., Aristimuno, L., et al. (2006) *Mycobacterium tuberculosis* complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology. BMC Microbiol. 6, 23.

[5] Brussow, H., Desiere, F. (2001) Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. Mol. Microbiol. 39, 213–222.

[6] Campbell, A. (1994) Comparative molecular biology of lambdoid phages. Annu. Rev. Microbiol. 48, 193–222.

[7] Casjens, S., Hatfull, G.F., Hendrix, R. (1992) Evolution of dsDNA tailed-bacteriophage genomes. Semin. Virol. 3, 383–397.

[8] Casjens, S., Hendrix, R. (1974) Comments on the arrangement of the morphogenetic genes of bacteriophage lambda. J. Mol. Biol. 90, 20–25.

[9] Clark, A.J., Inwood, W., Cloutier, T., Dhillon, T.S. (2001) Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. J. Mol. Biol. 311, 657–679.

[10] Doke, S. (1960) Studies on mycobacteriophages and lysogenic mycobacteria. J. Kumamoto Med. Soc. 34, 1360–1373.

[11] Ford, M.E., Sarkis, G.J., Belanger, A.E., Hendrix, R.W., Hatfull, G.F. (1998) Genome structure of mycobacteriophage D29: implications for phage evolution. J. Mol. Biol. 279, 143–164.

[12] Froman, S., Will, D.W., Bogen, E. (1954) Bacteriophage active against *Mycobacterium tuberculosis* I. Isolation and activity. Am. J. Public Health 44, 1326–1333.

[13] Garcia, M., Pimentel, M., Moniz-Pereira, J. (2002) Expression of Mycobacteriophage Ms6 lysis genes is driven by two sigma(70)-like promoters and is dependent on a transcription termination signal present in the leader RNA. J. Bacteriol. 184, 3034–3043.

[14] Ghosh, P., Wasil, L.R., Hatfull, G.F. (2006) Control of Phage Bxb1 excision by a novel recombination directionality factor. PLoS Biol. 4, e186.

[15] Hatfull, G.F. (1994) Mycobacteriophage L5: a toolbox for tuberculosis. ASM News 60, 255–260.

[16] Hatfull, G.F. (2006). In: R. Calendar (Ed.), The Bacteriophages (pp. 602–620). New York: Oxford University Press.

[17] Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., et al. (2006) Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. 2, e92.

[18] Hatfull, G.F., Sarkis, G.J. (1993) DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. Mol. Microbiol. 7, 395–405.

[19] Hendrix, R.W. (2003) Bacteriophage genomics. Curr. Opin. Microbiol. 6, 506–511.

[20] Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., Casjens, S. (2000) The origins and ongoing evolution of viruses. Trends Microbiol. 8, 504–508.

[21] Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., Hatfull, G.F. (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. U.S.A. 96, 2192–2197.

[22] Hussain, T. (2007) Leprosy and tuberculosis: an insight-review. Crit. Rev. Microbiol. 33, 15–66.

[23] Jacobs Jr., W.R. (2000). In: G.F. Hatfull, & W.R. Jacobs Jr. (Eds.), Molecular Genetics of the Mycobacteria (pp. 1–16). Washington, DC: ASM Press.

[24] Jacobs Jr., W.R., Tuckman, M., Bloom, B.R. (1987) Introduction of foreign DNA into mycobacteria using a shuttle phasmid. Nature 327, 532–535.

[25] Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., Hendrix, R.W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J. Mol. Biol. 299, 27–51.

[26] Kim, A.I., Ghosh, P., Aaron, M.A., Bibb, L.A., Jain, S., Hatfull, G.F. (2003) Mycobacteriophage Bxb1 integrates into the *Mycobacterium smegmatis* groEL1 gene. Mol. Microbiol. 50, 463–473.

[27] Kwan, T., Liu, J., DuBow, M., Gros, P., Pelletier, J. (2005) The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. Proc. Natl. Acad. Sci. U.S.A. 102, 5174–5179.

[28] Kwan, T., Liu, J., Dubow, M., Gros, P., Pelletier, J. (2006) Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. J. Bacteriol. 188, 1184–1187.

[29] Levy, L., Ji, B. (2006) The mouse foot-pad technique for cultivation of *Mycobacterium leprae*. Lepr. Rev. 77, 5–24.

[30] Lewis, J.A., Hatfull, G.F. (2001) Control of directionality in integrase-mediated recombination: examination of recombination directionality factors (RDFs) including Xis and Cox proteins. Nucleic Acids Res. 29, 2205–2216.

[31] Morris, P., Marinelli, L.J., Jacobs-Sera, D., Hendrix, R.W., Hatfull, G.F. (2008) Genomic characterization of mycobacteriophage giles: evidence for phage acquisition of host DNA by illegitimate recombination. J. Bacteriol. 190, 2172–2182.

[32] Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., et al. (2003) Origins of highly mosaic mycobacteriophage genomes. Cell 113, 171–182.

[33] Pham, T.T., Jacobs-Sera, D., Pedulla, M.L., Hendrix, R.W., Hatfull, G.F. (2007) Comparative genomic analysis of mycobacteriophage Tweety: evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. Microbiology 153, 2711–2723.

[34] Sahu, K., Gupta, S.K., Ghosh, T.C., Sau, S. (2004) Synonymous codon usage analysis of the mycobacteriophage Bxz1 and its plating bacteria *M. smegmatis*: identification of highly and lowly expressed genes of Bxz1 and the possible function of its tRNA species. J. Biochem. Mol. Biol. 37, 487–492.

[35] Saunders, N.A. (1995) State of the art: typing *Mycobacterium tuberculosis*. J. Hosp. Infect. 29, 169–176.

[36] Scollard, D.M., Adams, L.B., Gillis, T.P., Krahenbuhl, J.L., Truman, R.W., Williams, D.L. (2006) The continuing challenges of leprosy. Clin. Microbiol. Rev. 19, 338–381.

[37] Sorek, R., Kunin, V., Hugenholtz, P. (2008) CRISPR – a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol. 6, 181–186.

[38] Susskind, M.M., Botstein, D. (1978) Molecular genetics of bacteriophage P22. Microbiol. Rev. 42, 385–413.

[39] Timme, T.L., Brennan, P.J. (1984) Induction of bacteriophage from members of the *Mycobacterium avium, Mycobacterium intracellulare, Mycobacterium scrofulaceum* serocomplex. J. Gen. Microbiol. 130, 2059–2066.

[40] van Kessel, J.C., Hatfull, G.F. (2007) Recombineering in *Mycobacterium tuberculosis*. Nat. Methods 4, 147–152.

[41] van Kessel, J.C., Hatfull, G.F. (2008) Efficient point mutagenesis in mycobacteria using single-stranded DNA recombineering: characterization of antimycobacterial drug targets. Mol. Microbiol. 67, 1094–1107.