3-2014

Cluster M Mycobacteriophages Bongo, PegLeg, and Rey with Unusually Large Repertoires of tRNA Isotypes

Welkin H. Pope

Kirk R. Anders

Madison Baird

Charles A. Bowman

Michelle M. Boyle

See next page for additional authors

Authors
Welkin H. Pope, Kirk R. Anders, Madison Baird, Charles A. Bowman, Michelle M. Boyle, Gregory W. Broussard, Tiffany Chow, Kari L. Clase, Shannon Cooper, Kathleen A. Cornely, Randall J. DeJong, Veronique A. Delesalle, Lisa Deng, David Dunbar, Nicholas P. Edgington, Christina M. Ferreira, Kathleen Weston Hafer, Grant A. Hartzog, J. Robert Hatherill, Lee E. Hughes, Khristina Ipapo, Greg P. Krukonis, Christopher G. Meier, Denise L. Monti, Matthew R. Olm, Shallee T. Page, Craig L. Peebles, Claire A. Rinehart, Michael R. Rubin, Daniel A. Russell, Erin R. Sanders, Morgan Schoer, Christopher D. Shaffer, James Wherley, Edwin Vazquez, Han Yuan, Daiyuan Zhang, Steven G. Cresawn, Deborah Jacobs-Sera, Roger W. Hendrix, and Graham F. Hatfull

JVI Journals.ASM.org

Cluster M Mycobacteriophages Bongo, PegLeg, and Rey with Unusually Large Repertoires of tRNA Isotypes

Welkin H. Pope,[a] Kirk R. Anders,[b] Madison Baird,[c] Charles A. Bowman,[a] Michelle M. Boyle,[a] Gregory W. Broussard,[a] Tiffany Chow,[d] Kari L. Clase,[e] Shannon Cooper,[d] Kathleen A. Cornely,[f] Randall J. DeJong,[g] Veronique A. Delesalle,[h] Lisa Deng,[c] David Dunbar,[i] Nicholas P. Edgington,[j] Christina M. Ferreira,[a] Kathleen Weston Hafer,[c] Grant A. Hartzog,[k] J. Robert Hatherill,[l] Lee E. Hughes,[m] Khristina Ipapo,[d] Greg P. Krukonis,[h] Christopher G. Meier,[a] Denise L. Monti,[n] Matthew R. Olm,[a] Shallee T. Page,[o] Craig L. Peebles,[a] Claire A. Rinehart,[p] Michael R. Rubin,[q] Daniel A. Russell,[a] Erin R. Sanders,[d] Morgan Schoer,[c] Christopher D. Shaffer,[c] James Wherley,[c] Edwin Vazquez,[q] Han Yuan,[c] Daiyuan Zhang,[l] Steven G. Cresawn,[r] Deborah Jacobs-Sera,[a] Roger W. Hendrix,[a] Graham F. Hatfull[a]

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA[a]; Department of Biology, Gonzaga University, Spokane, Washington, USA[b]; Department of Biology, Washington University in St. Louis, St. Louis, Missouri, USA[c]; Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, California, USA[d]; Departments of Technology Leadership and Innovation and Agricultural and Biological Engineering, Purdue University, West Lafayette, Indiana, USA[e]; Department of Biochemistry, Providence College, Providence, Rhode Island, USA[f]; Department of Biology, Calvin College, Grand Rapids, Michigan, USA[g]; Department of Biology, Gettysburg College, Gettysburg, Pennsylvania, USA[h]; Department of Biology, Cabrini College, Radnor, Pennsylvania, USA[i]; Department of Biology, Southern Connecticut State University, New Haven, Connecticut, USA[j]; Department of Molecular, Cell & Developmental Biology, University of California Santa Cruz, Santa Cruz, California, USA[k]; Department of Natural Sciences, Del Mar College, Corpus Christi, Texas, USA[l]; Department of Biological Sciences, University of North Texas, Denton, Texas, USA[m]; Department of Biology, University of Alabama at Birmingham, Birmingham, Alabama, USA[n]; Department of Environmental and Biological Sciences, University of Maine at Machias, Machias, Maine, USA[o]; Department of Biology, Western Kentucky University, Bowling Green, Kentucky, USA[p]; Department of Biology, University of Puerto Rico at Cayey, Cayey, Puerto Rico[q]; Department of Biology, James Madison University, Harrisonburg, Virginia, USA[r]

ABSTRACT

Genomic analysis of a large set of phages infecting the common host *Mycobacterium smegmatis* mc^{2}155 shows that they span considerable genetic diversity. There are more than 20 distinct types that lack nucleotide similarity with each other, and there is considerable diversity within most of the groups. Three newly isolated temperate mycobacteriophages, Bongo, PegLeg, and Rey, constitute a new group (cluster M), with the closely related phages Bongo and PegLeg forming subcluster M1 and the more distantly related Rey forming subcluster M2. The cluster M mycobacteriophages have siphoviral morphologies with unusually long tails, are homoimmune, and have larger than average genomes (80.2 to 83.7 kbp). They exhibit a variety of features not previously described in other mycobacteriophages, including noncanonical genome architectures and several unusual sets of conserved repeated sequences suggesting novel regulatory systems for both transcription and translation. In addition to containing transfer-messenger RNA and RtcB-like RNA ligase genes, their genomes encode 21 to 24 tRNA genes encompassing complete or nearly complete sets of isotypes. We predict that these tRNAs are used in late lytic growth, likely compensating for the degradation or inadequacy of host tRNAs. They may represent a complete set of tRNAs necessary for late lytic growth, especially when taken together with the apparent lack of codons in the same late genes that correspond to tRNAs that the genomes of the phages do not obviously encode.

IMPORTANCE

The bacteriophage population is vast, dynamic, and old and plays a central role in bacterial pathogenicity. We know surprisingly little about the genetic diversity of the phage population, although metagenomic and phage genome sequencing indicates that it is great. Probing the depth of genetic diversity of phages of a common host, *Mycobacterium smegmatis*, provides a higher resolution of the phage population and how it has evolved. Three new phages constituting a new cluster M further expand the diversity of the mycobacteriophages and introduce novel features. As such, they provide insights into phage genome architecture, virion structure, and gene regulation at the transcriptional and translational levels.

The bacteriophage population is large, dynamic, old, and genetically diverse (1). Over 1,000 phage genomes have been sequenced, and the majority are double-stranded DNA (dsDNA) tailed phages, classified morphologically in the order *Caudovirales*. The genomes of dsDNA tailed phages vary in length from ~20 kbp to over 500 kbp (2) and typically contain 20 to 30 genes encoding virion structure and assembly functions, genes coding for DNA and nucleotide metabolism, a lysis cassette, and regulatory systems. The genomes of temperate phages typically encode repressors, contain operators, and frequently include systems for phage genome integration. However, with the exception of the few well-studied phage prototypes, the majority of genes carried by phages are of unknown function (1).

Mycobacteriophages—viruses of mycobacterial hosts such as *Mycobacterium smegmatis* and *Mycobacterium tuberculosis*—represent one of the largest collections of phages with completely sequenced genomes that are known to infect a single common

Received 15 November 2013 **Accepted** 9 December 2013
Published ahead of print 11 December 2013
Editor: M. J. Imperiale
Address correspondence to Graham F. Hatfull, gfh@pitt.edu.

doi:10.1128/JVI.03363-13

TABLE 1 Features of cluster M mycobacteriophages

Phage	Subcluster	Length (bp)	GC content (%)[a]	No. of ORFs[b]	No. of tRNAs	No. of tmRNAs	School performing sequencing[c]	Isolation location	Ends[d]	GenBank accession no.
Bongo	M1	80,228	61.6	132	22	1	WUSTL	St. Louis, MO	cos	JN699628
PegLeg	M1	80,955	61.5	138	21	1	UCLA	Los Angeles, CA	cos	KC900379
Rey	M2	83,724	60.9	152	24	1	Pitt	Houston, TX	cos	JF937105

[a] The average mycobacteriophage GC content is 64%.
[b] ORFs, open reading frames.
[c] WUSTL, Washington University at St. Louis; UCLA, University of California, Los Angeles; Pitt, University of Pittsburgh.
[d] All three phage genomes have cphesive (cos) ends.

host strain (*M. smegmatis* mc^{2}155) (3, 4). The 285 mycobacteriophage genomes currently available in GenBank are all dsDNA tailed phages classified morphologically as either siphoviral or myoviral and encompass substantial sequence diversity (4–6). Comparison at the nucleotide sequence level reveals groups of genomes that are more closely related to each other than to others, and these are referred to as clusters; members of a cluster typically share nucleotide sequence similarity that spans more than 50% of their genome lengths and have closely related gene contents (7). Currently, 15 clusters (cluster A to cluster O) as well as nine singleton genomes, each of which has no close relatives, have been described (6). These phages have been important for developing genetic approaches for the mycobacteria and tuberculosis diagnosis (4, 8–10) and more generally for genome engineering in heterologous systems (11, 12).

Although the degree of sequence diversity among the mycobacteriophages is considerable, they also have a broad array of notable features. For example, the cluster C phages and the singleton phage Wildcat have large sets of tRNA genes, whereas other phages have either no tRNA genes or perhaps one or two (13, 14). In the phages with siphoviral morphologies (all phages except those of cluster C), the 25 to 30 virion structure and assembly genes have a shared synteny. The genes either can be tightly packed and occupy minimal genome space (e.g., 23 kbp in cluster G phages [15]) or can be interspersed with apparently nonstructure genes expanding the operon to over 35 kbp (e.g., cluster J [5]). In phage Marvin, several of the tail genes are displaced and relocated about 20 kbp from their canonical position (16). Cluster A phages have a complex immunity system in which the repressor binds to multiple stoperator sites located throughout the genome (14, 17), and in the cluster G phages, establishment of immunity is unusually dependent on phage integration (18). Some phage genomes have multiple repeated sequences that are not obviously associated with immunity, including the translation start-associated sequences (SASs) in the cluster K phages, a subset of which also contains an upstream extended start-associated sequence (ESAS) (19). The origins of all this diversity are not clear, but it has been suggested that it reflects the ability of the viruses to relatively rapidly switch from one host to another. This ability then enables a speedy migration across the bacterial landscape, provided that a diverse set of relatively closely related bacterial cells is present in that environment (20). Viral genomes are architecturally mosaic, and gene acquisition by horizontal exchange provides a mechanism for rapid adaptation to new host environments (13, 20, 21).

We describe here three mycobacteriophages that are not closely related to other phages and constitute a new cluster, cluster M. The genomes of phages Bongo, PegLeg, and Rey are longer than the average mycobacteriophage genome (80.2 to 83.7 kbp) and have several features of interest, including unusual and complex sets of conserved repeated sequences and intriguing suites of tRNA genes with potential regulatory roles.

MATERIALS AND METHODS

Isolation, genome sequencing, and bioinformatic analysis. Mycobacteriophage Rey was isolated by direct plating of a soil sample collected in Houston, TX, on lawns of *M. smegmatis* mc^{2}155; Bongo and PegLeg were isolated by enrichment with *M. smegmatis* mc^{2}155 from soil samples collected in St. Louis, MO, and Los Angeles, CA, respectively (Table 1). High-titer lysates were prepared and DNA extracted, and the genomes were sequenced using 454 pyrosequencing. Bongo, PegLeg, and Rey were sequenced at the Genome Institute of Washington University in St. Louis, the University of California, Los Angeles, Genotyping and Sequencing Core, and the University of Pittsburgh Genomics and Proteomics Core Laboratories, respectively. Genomes were assembled using the NewBler and Consed programs (22), and the assemblies were reviewed for completion at the Pittsburgh Bacteriophage Institute. Genomes were annotated using the DNA Master (http://cobamide2.bio.pitt.edu/), GLIMMER (23), GeneMark (24), BLAST, HHPred (25), and Phamerator (Mycobacteriophage_285 database) (26) programs. Dot plots were constructed using the Gepard program (27), and further bioinformatic analyses were performed using the DNA Master, FindTerm (Softberry), Splitstree (28), GLAM2, GLAM2SCAN, and MEME (29) programs. tRNAs were identified using the Aragorn (v1.2.36) program with default settings (30) or by the tRNAscan-SE (v1.23) program (31, 32) with default settings, except for bacterial source, relaxed scan modes, and s Cove score of >2. Following comparative analyses, revisions were made to the tRNA genes that included a deletion of mycobacteriophage Rey gene *113*.

Lysogeny and immunity assays of cluster M phages. Serial dilutions of high-titer lysates of Bongo and Rey were spotted onto lawns of mc^{2}155, and turbid infected areas were picked and streaked for single colonies.

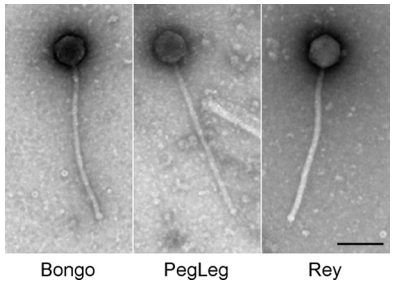


FIG 1 Virion morphologies of mycobacteriophages Bongo, PegLeg, and Rey. Bongo, PegLeg, and Rey samples were put on carbon-coated 400-mesh copper grids and stained with 1% uranyl acetate. Phages were visualized by transmission electron microscopy. All three phages have an isometric icosahedral head and flexible tail, with heads 64 nm in diameter and tails 348 nm long. Bar, 100 nm.

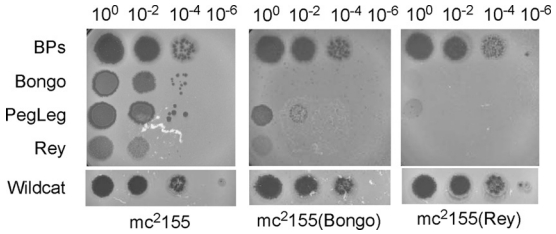
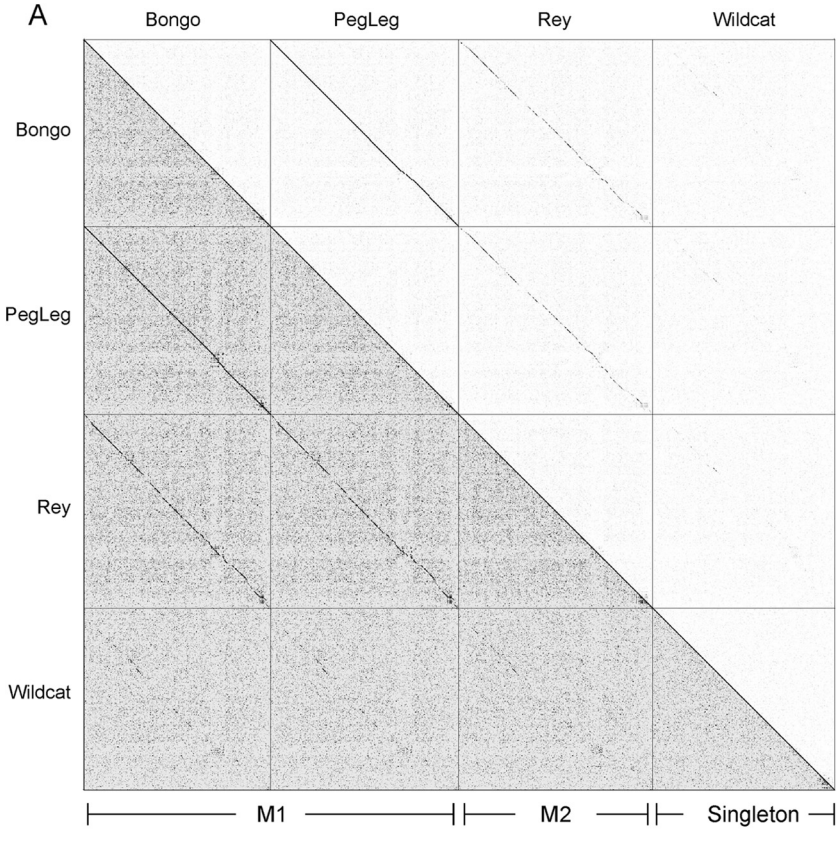


FIG 2 Immune specificities of cluster M lysogens. Lysogens of Bongo and Rey were isolated, and the three cluster M phages and Wildcat were tested for their ability to infect the lysogens through serial dilution and spotting of the lysates on the lysogen lawns. The cluster M phages are homoimmune but do not confer immunity to Wildcat. BPs, a cluster G phage, is included as a positive control.

Colonies were patched onto Middlebrook 7H10 plates and overlaid with soft agar containing mc^{2}155. Putative lysogens were identified by clearing of the overlaid lawn—reflecting phage release—and grown in liquid 7H9 medium. Lysogeny was confirmed by immunity and phage release.

Electron microscopy of cluster M phages. High-titer lysates of cluster M phages were serially diluted and spotted on a Middlebrook 7H10 plate overlaid with soft agar and *M. smegmatis* mc^{2}155; plates were incubated for 24 h at 37°C. Spots with confluent plaques were gently washed with 10 μl of phage buffer several times by pipetting up and down to recover phage particles. The liquid phage sample was then diluted 1:2 with phage buffer, and 5 μl was placed on carbon-Formvar-coated glow-discharged copper grids. The sample was allowed to stand for 30 s, rinsed with double-distilled H_2O, stained with 1% uranyl acetate, and visualized with an FEI Morgagni transmission electron microscope (TEM).

Mass spectrometry of Rey virions. A high-titer lysate of Rey (10^{10} PFU/ml) was precipitated overnight with 10% polyethylene glycol 8000, collected by centrifugation (5,500 × *g*, 10 min), and resuspended in phage buffer for 30 min. The suspension was mixed with CsCl (~0.85 g/ml) and sealed in a heat-sealed ultracentrifuge tube (Beckman). The sample was spun for 16 h at 85,000 × *g* in a Ti70.1 rotor, and the visible phage band was collected and dialyzed against phage buffer. Phage particles were pelleted in a microcentrifuge by spinning at top speed for 30 min. The phage pellet was resuspended in 75 μl of 20 mM dithiothreitol, 2 μl of 0.5 M EDTA was added, and the particles were heated to 70°C. The sample was chilled on ice for 10 min, sonicated twice for 10 s, mixed with SDS sample buffer, and loaded onto a 12% SDS-polyacrylamide gel. The proteins were electrophoresed until they just barely entered the resolving portion of the gel, such that all proteins were contained within a narrow region. The gel



B

	Bongo	PegLeg	Rey	Wildcat
Bongo	1	0.995	0.739	0.609
PegLeg	0.995	1	0.738	0.610
Rey	0.739	0.738	1	0.611
Wildcat	0.609	0.610	0.611	1

FIG 3 Nucleotide sequence comparisons of cluster M phages. (A) Dot plot comparison of mycobacteriophage genomes. Bongo, PegLeg, Rey, and Wildcat nucleotide sequences were catenated and compared using the program Gepard. Bottom left and top right triangles, less and more stringent outputs, respectively. (B) Average nucleotide identities. The nucleotide sequences of all cluster M genomes and Wildcat were compared to each other using DNA Master's average nucleotide identity algorithm.

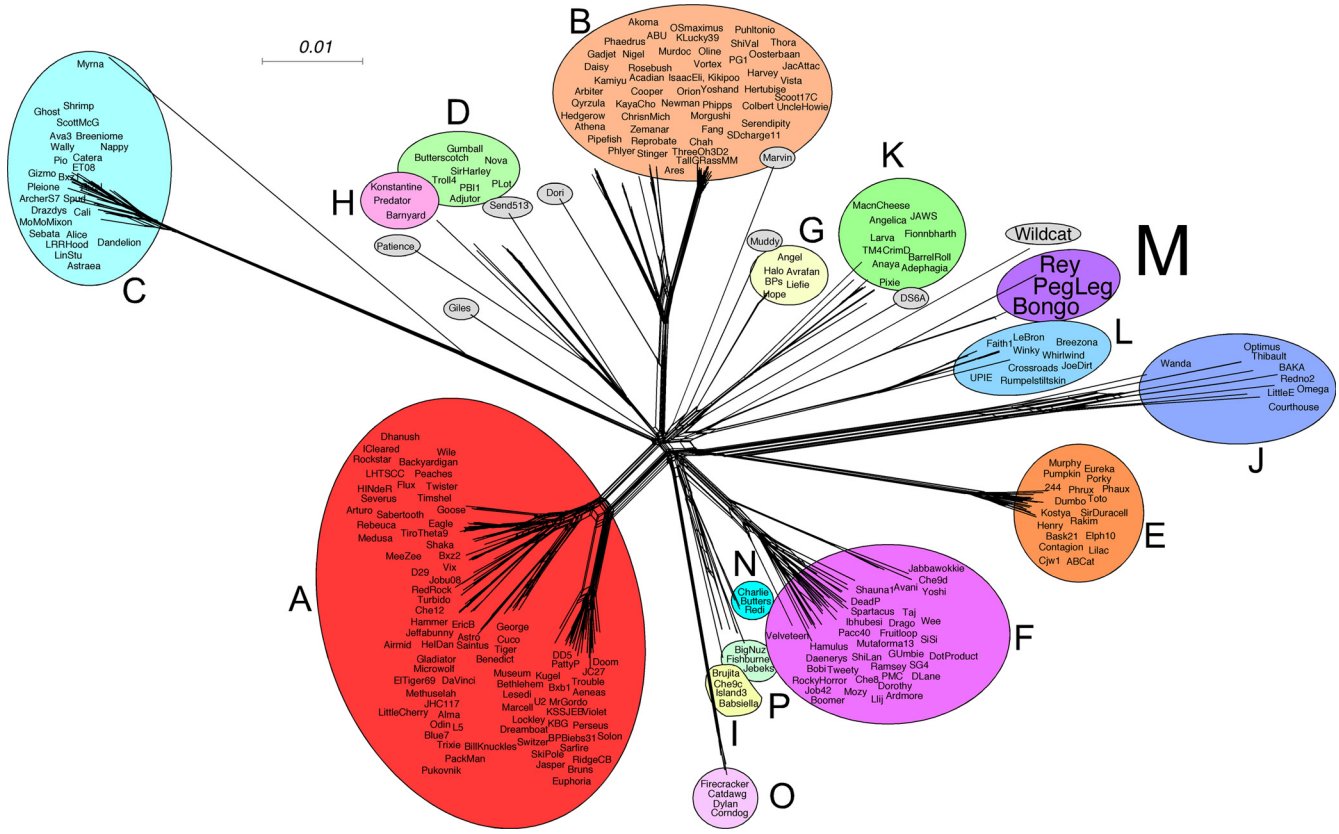


FIG 4 Splits Tree representation of the gene content of 285 sequenced and annotated mycobacteriophages. The 32,802 genes of the 285 genomes were compared to each pairwise by use of the BLASTP and ClustalW programs. Similar genes were sorted into 3,435 phams using the program Phamerator (26). Genomes were determined to have the same gene if they had members in the same pham. The overall gene content of each genome is shown here using the program Splits Tree.

was stained with Coomassie blue, the protein band was excised, and the proteins were subjected to in-gel trypsin digestion. Tryptic fragments were eluted from the gel, run on a reverse-phase high-pressure liquid chromatography (HPLC) column, and subjected to electrospray and tandem mass spectrometry (MS/MS) in an LTQ ion trap MS. Mass spectra were analyzed using the Scaffold (v4) program.

Nucleotide sequence accession numbers. Revised genome annotations were submitted to GenBank (accession numbers JF937105, JN699628, and KC900379).

RESULTS

Bongo, PegLeg, and Rey constitute members of mycobacteriophage cluster M. Mycobacteriophages Bongo, PegLeg, and Rey were isolated in St. Louis, MO; Los Angeles, CA; and Houston, TX, respectively (6). All three have siphoviral morphologies (Fig. 1). All three cluster M phages make turbid plaques on *M. smegmatis* mc^{2}155, and we recovered lysogens from Bongo and Rey infections. These lysogens were tested for immunity to infection by all three cluster M phages, as well as to Wildcat and BPs, a member of cluster G (Fig. 2). Bongo and Rey are likely part of the same immunity group but do not confer immunity to either BPs or Wildcat.

The three cluster M genomes vary in length from 80,228 bp to 83,724 bp (Table 1), lengths that are somewhat longer than the average mycobacteriophage genome length of 70 kbp (4). Dot plot analysis shows that they share substantial DNA sequence similarity spanning greater than 50% of their genome lengths (Fig. 3A), and they share high levels of average nucleotide sequence identity (Fig. 3B). A gene content network analysis clearly shows the similarity between Bongo, PegLeg, and Rey (Fig. 4), and alignment of the genome maps indicates closely related architectures (Fig. 5). Taken together, phages Bongo, PegLeg, and Rey are clearly more closely related to each other than to other mycobacteriophages, and we propose that these constitute cluster M. Accession numbers are shown in Table 1.

Several features of the singleton phage Wildcat are shared with the cluster M phages (e.g., tRNA arrays, gene content, and genome architecture; see below), and dot plot analysis showed weak but identifiable sequence similarity (Fig. 3). However, alignment of any of the cluster M phages with Wildcat using Align Two sequences with BLASTN showed that the similarity spanned less than 50% of their genome lengths, and we conclude that Wildcat is insufficiently closely related to Bongo, PegLeg, and Rey to warrant inclusion in cluster M.

Although the three cluster M genomes are clearly related, Bongo and PegLeg are considerably more closely related to each other than either is to Rey, as illustrated by the branching pattern seen in the Splits Tree representation of their gene content relationships (Fig. 4), by dot plot analysis (Fig. 3A), and by the differences in the average nucleotide sequence identity values (Fig. 3B). We therefore propose that cluster M is divided into two subclusters, with Bongo and PegLeg forming subcluster M1 and Rey representing subcluster M2 (Table 1). Alignment of the genome maps further supports this division (Fig. 5).

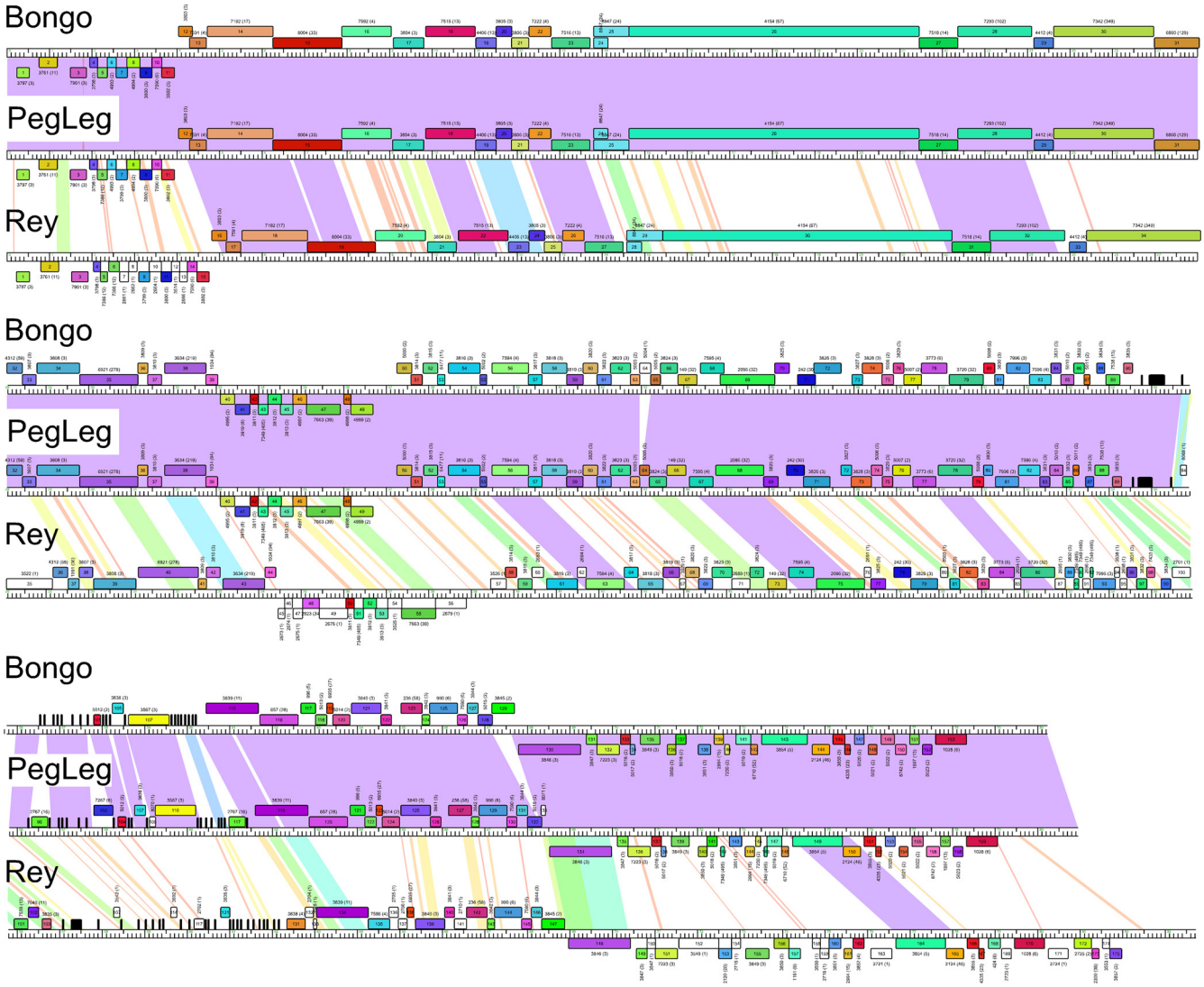


FIG 5 Comparison of cluster M phage genomes. The three cluster M genomes are shown aligned with each other and drawn in three tiers. Maps were generated using Phamerator (26) and have the following features. The center ruler for each phage shows the genome length (in kilobases), and rightward- and leftward-transcribed genes are shown as boxes above and below the genome, respectively. Gene boxes are colored according to phamily membership, with each pham number listed above the gene and the number of pham members given in parentheses. Spectrum-colored shading between the genomes displays nucleotide sequence similarity determined by BLASTN, with the most similar shown in violet and the least similar shown in red, with a minimal cutoff E value of 10^{-4}. This figure illustrates the overall relationships between these genomes, and more detailed maps of each of them are shown in Fig. 6 to 8.

Unusual genome architecture of cluster M phages. There are several noncanonical features of the cluster M phages that distinguish them from other mycobacteriophage genomes that have been described. First, although the organization of the virion structure and assembly genes is typical of other phages with siphoviral morphologies—ordered as terminase, portal, protease, scaffolding, major capsid subunit, head-tail connector proteins, major tail subunit, tail assembly chaperones, tapemeasure protein, and minor tail proteins—the terminase is separated from the left cohesive end by 4 to 5 kbp DNA containing 10 to 15 small, leftward-transcribed genes (Fig. 6 to 8). More than 30% of the Rey genes are different from those in Bongo and PegLeg, and most of the genes are of unknown function (Fig. 5). The exceptions are the gp3 genes of Bongo, PegLeg, and Rey that encode Lsr2-like proteins, and although Lsr2 proteins have been described in mycobacteriophages within clusters E, J, and L, they are typically located within right-arm genes, and their specific roles are ill defined. In *Mycobacterium* spp., Lsr2 proteins are small DNA-bridging proteins that may play a role in DNA compaction. The lysis cassette is located immediately downstream of the virion structural genes, a common location for all the mycobacteriophages except those in cluster A (Fig. 5).

The integration cassette of cluster M phages is in an unusual location, positioned ~12 kbp (15% of the genome length) from the genome right ends (Fig. 5 to 8). Among the mycobacteriophages, the integration cassette is typically located near the center of the genome, irrespective of genome length. The greatest departure from the genome center reported previously is in phage Brujita, in which the integrase is positioned 39% of the genome length from the left end, although we note that there are other examples

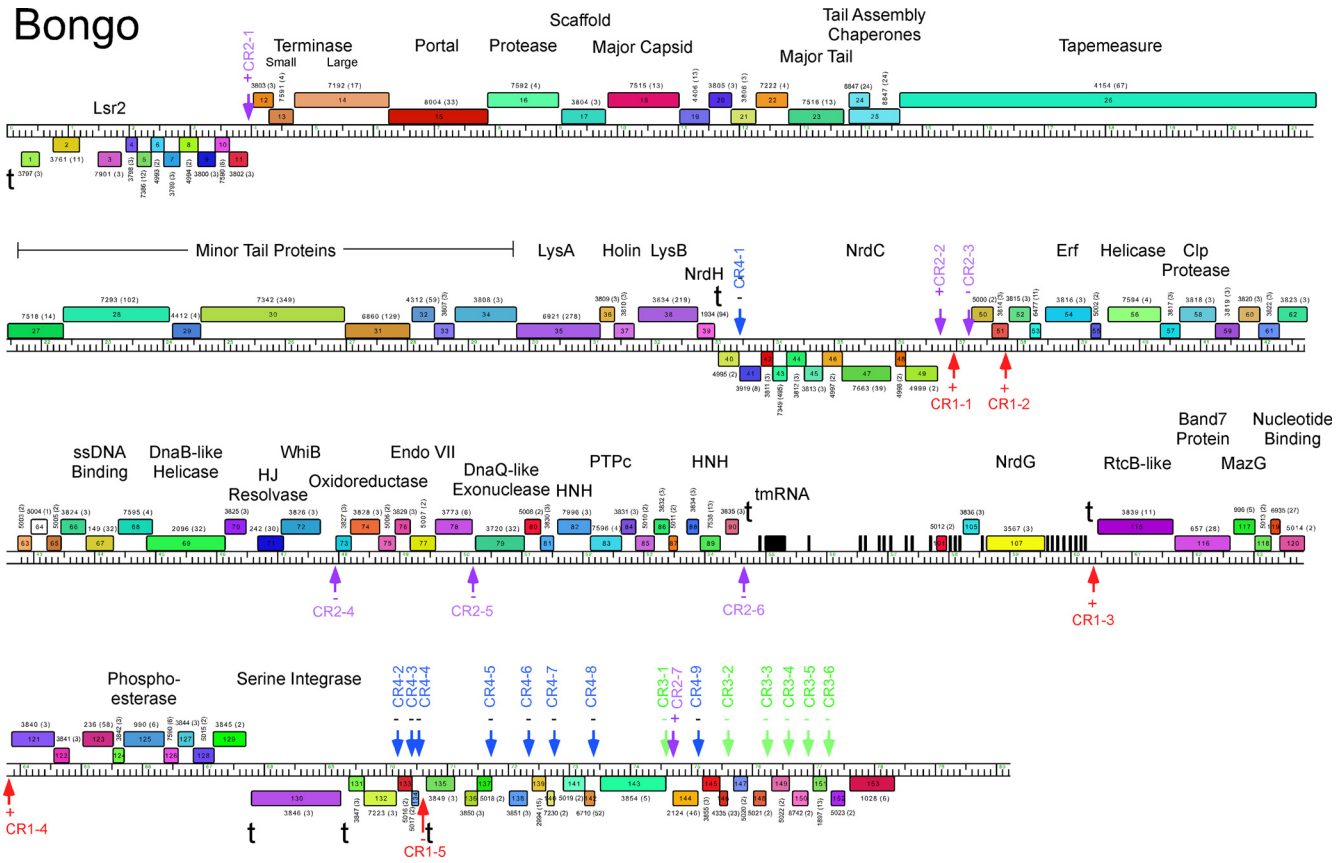


FIG 6 Genome map of mycobacteriophage Bongo. The annotated map of the Bongo genome is shown, with features displayed as described in the legend to Fig. 5. Putative functions are shown above genes, and terminators (t) are shown on either the top or bottom strand. The locations of conserved repeats (CR1, CR2, CR3, CR4) are indicated. HJ resolvase, Holliday junction resolvase.

of nonmycobacteriophages where the integrase is not centrally located, including *Streptomyces* phages ϕC31 (33) and R4 and their respective relatives (34). Although many of the genes surrounding the integration cassette in the cluster M phages are of unknown function, several genes involved in DNA and RNA metabolism can be identified, and there is an unusually large cluster of tRNA genes. Each of these features is discussed in further detail below.

Virion structure and assembly genes. The 22 virion structure and assembly genes are tightly linked in a rightward-transcribed operon in the left arms of the genomes. Although some mycobacteriophage genomes—especially those in cluster J (5)—are replete with insertions of nonstructural genes throughout the structural operon, this is not observed in the cluster M phages. The sole possible exception is the first rightward-transcribed gene (e.g., Bongo *12*; Fig. 6); there are few bioinformatic clues as to its function or whether it is expressed, but it is not uncommon to find HNH genes in similar locations in other genomes. A striking feature of the operon is the exceptionally long tapemeasure gene (*tmp*; 6.8 kbp), the longest of any in mycobacteriophages described, reflecting the long tails of the cluster M phages (Fig. 1). The correlation between *tmp* gene length and tail length (one codon/1.5 Å) reflects that reported for other siphoviral phages (13). Interestingly, the Tmp protein of the subcluster M2 phage Rey has substantially diverged from the Tmp proteins of Bongo and PegLeg and shares only 47% amino acid identity, in contrast to the two tail genes to its right, which share 89% and 86% amino acid identities, respectively, with those of the subcluster M1 phages. However, all three have features that are common to tapemeasure proteins, with a high proportion of glycine plus alanine residues (>20%), strongly predicted coiled-coil domains near their N termini, and a predicted peptidoglycan-hydrolyzing motif in the C-terminal half of the protein. This transglycosylase motif belongs to Tmp motif 1 described previously (13), although it is a relative far distant from those in other mycobacteriophages.

The cluster M major capsid subunits are predicted by HHPred to contain an HK97-like fold and, in addition to being related to the Wildcat capsid subunit (47% amino acid identity), have sequence similarity to the cluster L phages, such as LeBron and Rumpelstiltskin (37% amino acid identity). Like Wildcat and the cluster L phages (4, 14), the cluster M major capsid subunit and major tail proteins contain predicted Ig-like domains in their C-terminal ~100 residues, and they are weakly related to each other (~30% amino acid identity), a relationship also seen in Bxb1 and related phages (35). The Ig-like protein domains are widespread in phage proteins and have been proposed to play accessory roles in phage infection by weakly interacting with carbohydrates on the bacterial cell surface and by allowing the phages to bind to glycans from mucosal membranes of metazoans, allowing the phages access to bacterial inhabitants of those layers (36, 37).

Following the tapemeasure protein gene are eight predicted minor tail proteins (Fig. 6 to 8). The first two of these (e.g., Bongo

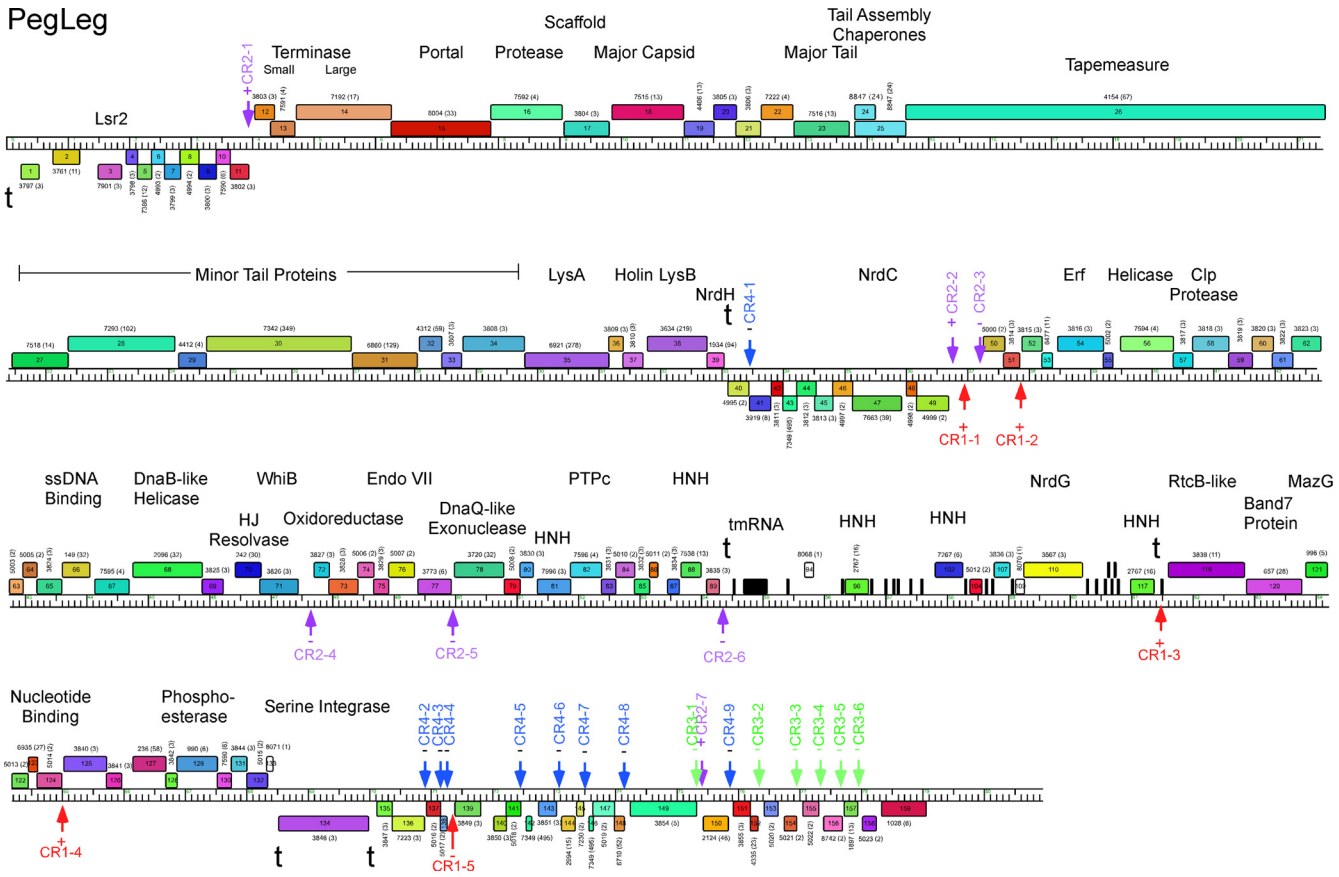


FIG 7 Genome map of mycobacteriophage PegLeg. The annotated map of the PegLeg genome is shown, with features displayed as described in the legend to Fig. 5. Putative functions are shown above genes, and terminators (t) are shown on either the top or bottom strand. The locations of conserved repeats (CR1, CR2, CR3, CR4) are indicated.

gene *27* and Bongo gene *28*) are closely related in all three cluster M phages (>80% amino acid identity), but the others are more distantly related between the subcluster M1 and M2 phages. For example, Rey gp35 and gp36 have little recognizable sequence similarity to their counterparts in Bongo (gp31 and gp32, respectively). Rey gp35 is glycine rich (>16%)—a not uncommon feature of phage tail fiber proteins—and is more closely related to cluster L tail proteins (e.g., LeBron gp21; 27% amino acid identity) than to the cluster M1 phages. The minor tail protein gene cluster extends through to Bongo gene *34*, PegLeg gene *34*, and Rey gene *39*, and the last of these is related to the gene for Marvin gp57, which has been shown to be a virion component (16). In addition, Rey gp39 was identified as a virion component by mass spectrometry (see below; Table 2).

The protein composition of Rey was analyzed using tandem mass spectrometry. Rey virions were purified by CsCl density gradient ultracentrifugation and then electrophoresed into a single band in an SDS-polyacrylamide gel. The particle proteins from the gel slice were digested with trypsin, and the resulting tryptic peptides were concentrated using reverse-phase HPLC, followed by loading into an ion-trap mass spectrometer. Peptides were further fragmented and analyzed for the mass and charge of the composite pieces. Spectra were compared to those in a database generated from the coding sequences of the Rey genome using the program Scaffold (v4) (Table 2).

The most abundant Rey virion peptides correspond to the major capsid protein and the major tail subunit, and other expected virion proteins were identified, including the tapemeasure (gp30), portal (gp19), minor tail component (gp31, gp39), and head-to-tail connector (gp23, gp24, and gp26) proteins. However, several predicted virion proteins were not identified with high confidence, including minor tail proteins gp32, gp33, gp34, gp35, gp36, gp37, and gp38. Peptides for most of these were identified in the mass spectrometry data, but at lower confidence levels, perhaps in part reflecting a lower abundance of these proteins in the particles or a lack of suitable trypsin cleavage sites. A peptide from the large terminase subunit was also detected, a finding which was not expected and could indicate contamination of the sample with incompletely packaged particles. Two independent mass spectrometry runs generated similar profiles. Surprisingly, three proteins encoded by right-arm genes were identified as virion components with high confidence, an Erf-like recombinase (gp61), a single-stranded DNA (ssDNA) binding protein (SSB; gp73), and a ClpP-like protease (gp65) (Table 2). While we cannot rule out the possibility that these are simply contaminants of the initial phage sample, examination of the CsCl-banded phage sample by TEM showed no recognizable host or viral subparticles other than phage virions. The prevalence of the Erf peptides in the mass spectrometry data strongly suggests that the Erf protein in particular is indeed virion associated.

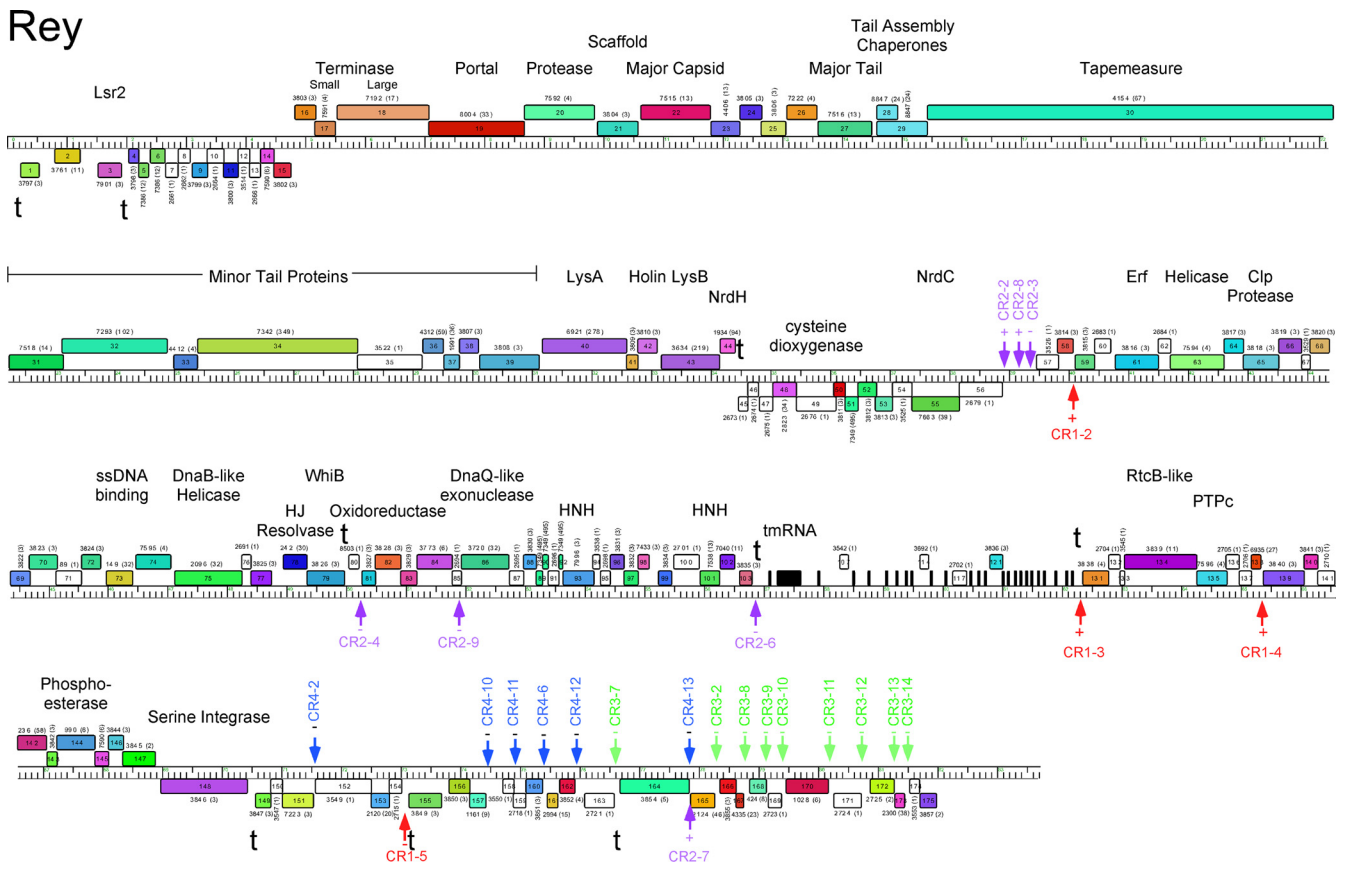


FIG 8 Genome map of mycobacteriophage Rey. The annotated map of the Rey genome is shown, with features displayed as described in the legend to Fig. 5. Putative functions are shown above genes and terminators (t) are shown on either the top or bottom strand. The locations of conserved repeats (CR1, CR2, CR3, CR4) are indicated.

Lysis cassette. Immediately downstream of the virion structural protein genes is the lysis cassette, which includes both lysin A and lysin B separated by two small genes. We propose that the second of these smaller genes (Bongo gp37, PegLeg gp37, and Rey gp42) is the holin, with each having two predicted transmembrane domains, although they are not related at the sequence level to any other mycobacteriophage proteins. The first small protein (e.g., Bongo gp36, 78 amino acids) contains an N-terminal predicted transmembrane domain, and its role is unclear, although it could perform a chaperone-like role in lysis, as proposed for gp1 of phage Ms6 (38). The lysin A proteins have domain structures corresponding to organization A (Org-A) (39), although the only other similar mycobacteriophage lysin is that encoded by the singleton phage Patience (gp41; 36% amino acid identity). The cluster M lysin B proteins are also distantly related to other lysin B proteins, with the closest non-cluster M relative being DS6A (42% amino acid identity) (6). The last gene in the rightward operon is a predicted NrdH-like glutaredoxin, and although related proteins are quite common in the mycobacteriophages, they are not typically found in this location (with the exception of the cluster J phages). However, it is unlikely that they play a role in cluster M phage lysis.

Immunity functions and the integration cassette. Although the cluster M phages are temperate and form stable lysogens, the genes responsible for repressor function cannot be easily identified. In other phages, it is common for the repressor to be divergently transcribed from an adjacent Cro-like protein (40), and this is seen in mycobacteriophages such as BPs (18) and Giles (41), but not in others, such as L5 and its relatives (14, 42). In the cluster M phages, there are only two regions with divergent transcription (e.g., Bongo genes *11* and *12* and genes *49* and *50*; Fig. 6), but they contain no proteins with predicted DNA binding activity. However, we note that there are examples of other mycobacteriophage repressors that also do not contain predicted DNA binding motifs (41). In some phages, the repressor genes are located near the integration functions (18), but none of the cluster M genes near the integrase has repressor-like features.

As noted above, each of the cluster M phages encodes an integrase of the serine recombinase family. There are many other similar mycobacteriophage integrases, although they are almost exclusively found in cluster A phages. One exception is that encoded by the cluster K phage Larva (6) and is the most closely related to the cluster M integrases (28% amino acid identity). Attachment sites for serine integrases cannot be readily predicted because of the lack of extended sequence similarity between *attP* and *attB* sites (43), although regions of imperfectly conserved inverted symmetry immediately to the right of the integrase genes are candidates for providing *attP* function. The intergenic regions to the right of the *int* genes are only moderately conserved between Rey and the subcluster M1 phages, raising the question as to whether they use the same *attB* site. Predicting the location of a putative recombination directionality factor (RDF) is complicated by the

TABLE 2 Rey particle proteins identified by tandem MS[f]

Gene, identified protein ($n = 13$)	Molecular mass (kDa)	Protein identification probability[a] (%)	Exclusive unique spectrum counts[b]	Unique peptide counts[c]	Total spectrum counts[d]	Protein coverage (%)[e]
Gene *22*, major capsid	42	100	44	26	379	80
Gene *27*, major tail	32	100	22	13	74	47
Gene *61*, Erf	25	100	9	7	25	39
Gene *23*, head-to-tail connector	18	100	6	6	15	42
Gene *19*, portal	60	100	9	9	14	14
Gene *73*, ssDNA binding	16	100	4	4	7	22
Gene *31*, minor tail	34	100	4	4	6	15
Gene *39*, minor tail	36	100	4	3	5	7.60
Gene *24*, head-to-tail connector	13	100	4	4	5	32
Gene *65*, ClpP-like protease	23	100	3	3	5	19
Gene *30*, tapemeasure	236	100	3	3	4	1.20
Gene *26*, head-to-tail connector	19	100	1	1	2	11
Gene *18*, terminase	56	99	1	1	2	1.20

[a] Protein identification probability was calculated by the Scaffold (v4) program on the basis of the recovered spectra matched to those in a database containing only predicted coding sequences in Rey.
[b] Exclusive unique spectrum counts, spectra identified in the data as matching only one specific protein in the database.
[c] Unique peptide counts, number of peptides with different sequences identified that matched a protein in the database.
[d] Total spectrum counts, all spectra from all peptide fragments generated that matched a protein in the database, including repeated matches.
[e] Protein coverage, percentage of identified peptides matched to the total length of the predicted protein.
[f] Peptides generated from an in-gel tryptic digestion of purified Rey particles were analyzed by tandem MS/MS, and the spectra were matched to those in a database generated from predicted Rey proteins using Scaffold (v4). Proteins were considered positively identified if they contained at least one unique peptide, with a peptide threshold match of 95% (as determined by the scoring algorithm of Scaffold [v4]) and a protein identification probability of 99%. Proteins were sorted by total spectrum counts, with all proteins having at least two spectra. Use of these criteria led to the inclusion of the terminase and a head-to-tail connector protein as the penultimate match.

variety of different proteins that can perform this function (44–46). However, we note that located near the integrases are genes encoding phosphoesterases (Bongo gp125, PegLeg gp129, Rey gp144), reminiscent of the RDF encoded by the gene for Bxb1 (gp47) (45), which is also predicted to function as a phosphoesterase. However, the phosphoesterase activity of Bxb1 gp47 is not required for its RDF function (47), and as Bxb1 gp47 and the cluster M phosphoesterases are unrelated at the sequence level, an RDF function cannot be confidently predicted.

Other nonstructural genes. In the rightmost ~48 kbp of the cluster M genomes, the genes appear to be organized into three large groups: a leftward-transcribed set of 10 to 12 genes (Fig. 6 to 8) immediately adjacent to the virion genes, a rightward-transcribed group of ~80 genes that includes many tRNA genes (discussed below), and a leftward-transcribed set of ~25 genes at the right ends of the genomes. The functions of less than 25% of these genes can be predicted using BLASTP and HHPred searches. Many of these are associated with nucleic acid metabolism and include DNA helicases, a DnaQ-like exonuclease, an ssDNA binding protein, an Erf-like recombinase, Holliday junction resolvase, nucleotide binding proteins, the putative RNA ligase component RtcB (48), and HNH endonucleases. All three genomes also encode proteins containing putative protein tyrosine phosphatase catalytic domain (PTPc) kinase/phosphatase domains, although the genes are located differently in the three genomes (Bongo gene *83*, PegLeg gene *82*, and Rey gene *135*). Additional predicted functions include a Band-7 protein, MazG, NrdC, NrdG, a ClpP-like protease, an oxidoreductase, and a WhiB-like regulator. Gene *152* in Rey has no counterpart in the subcluster M1 phages but is predicted to have a ParB domain-like fold and may play a regulatory role. The specific roles of these genes and their products are not known. However, as noted above, mass spectrometry suggests that the Erf protein and perhaps the SSB protein are virion associated. Inclusion of these proteins in capsids might make sense if they are required for recircularization of terminally redundant genome ends like Erf is in P22 (49), but genomic analysis shows that these phages have cohesive termini with 11-bp ssDNA extensions at the 5′ end (Table 1), and such an involvement of Erf is unexpected.

Transcription signals. Mycobacteriophage promoters are ill defined, and although some phages—including L5 and BPs and their relatives—use SigA-like promoters (18, 50), others apparently do not (41). A search for SigA-like promoters yields few strong candidates in any of the cluster M genes that are consistent with the gene organization, although we cannot exclude the possibility that regions with weaker matches for the SigA consensus could be functionally important. Nonetheless, we predict that there are likely to be at least five promoters for expression of the five apparent operons. Factor-independent transcription terminators can be confidently predicted, and we have identified six in Bongo and PegLeg (Fig. 6 and 7) and eight in Rey (Fig. 8). The locations in common to all three phages are at the extreme left ends of the genomes following the leftward-transcribed operon, at the end of the rightward-transcribed operons containing the structural genes and lysis cassette, immediately before and immediately following the groups of tRNA genes, and between *int* and the upstream gene (Fig. 6 to 8). Bongo and PegLeg also have a terminator immediately downstream of *int* (Fig. 6 and 7), and Rey has one following the *whiB* gene (gene *79*) and two within the leftward-transcribed genes at the right end of the genome (Fig. 8). Identification of the terminator upstream of the tRNA genes suggests that there is likely to be an additional promoter located between the terminator and tRNA genes.

Conserved repeated sequences. There are at least four short conserved repeated sequences in the cluster M phages that could play regulatory roles. The first—which we refer to as conserved repeat 1 (CR1)—is the 17-bp motif 5′-CTGACCTGCGATT

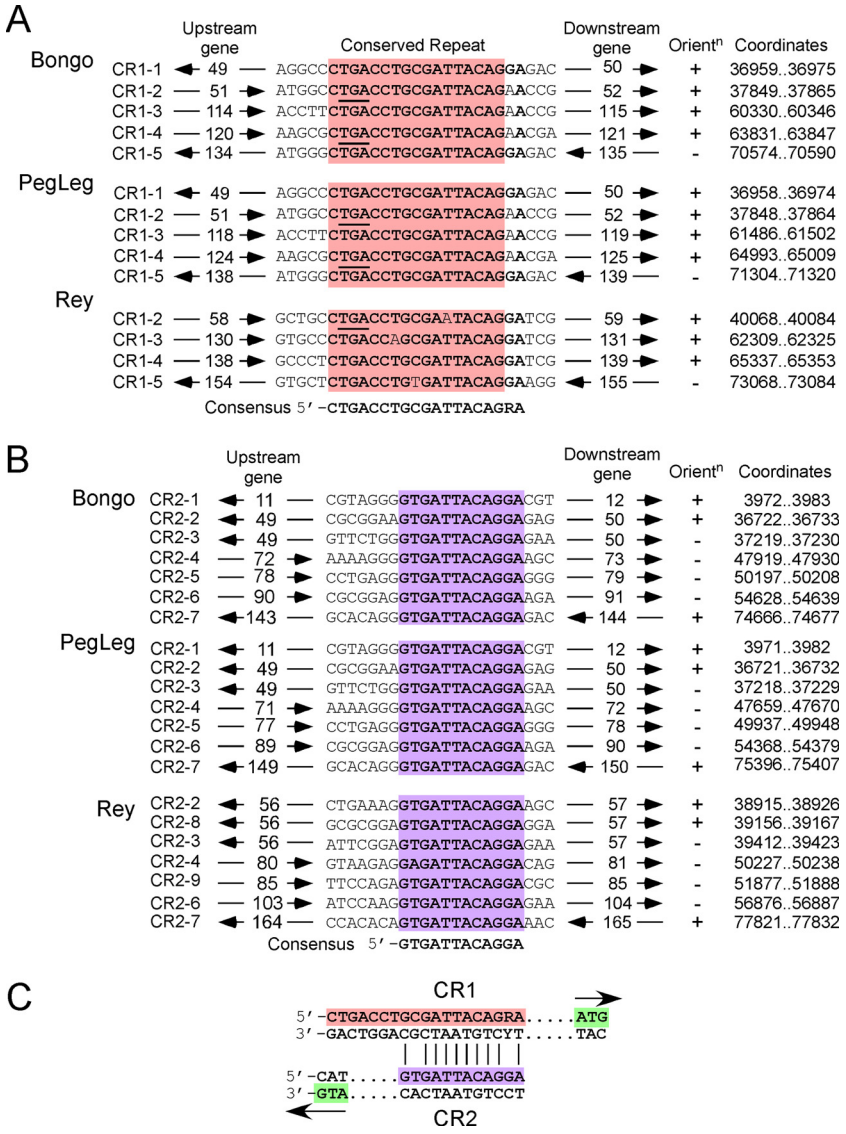


FIG 9 Conserved repeat sequences CR1 and CR2 in cluster M phages. Conserved repeated sequences CR1 (A) and CR2 (B) in Bongo, PegLeg, and Rey are shown, together with their flanking genes, the termination codon of the upstream gene (underlined), orientation (Orient^n), and coordinates. CR1 and CR2 sequences are boxed in red and purple, respectively, with their consensus sequences shown below. The directions of transcription of the flanking genes are indicated with arrows. For sites in the plus (+) orientation, the top-strand sequence is shown; for sites in the minus (−) orientation, the bottom-strand sequence is shown. (C) Alignment of the CR1 and CR2 consensus sequences shows that CR2 is a subset of CR1 but that whereas the CR1 consensus reflects the top strand relative to the direction of transcription of the downstream gene (with the start codon shown in green), CR2 is in the opposite orientation.

ACAG (Fig. 9A). In Bongo and PegLeg there are five identical copies of this asymmetric sequence (CR1-1 to CR1-5), four of which are oriented in one direction and one of which is oriented in the other. Four related sequences are present in Rey, although one shares the identical sequence, and the other three each have a different 1-base departure. The consensus could plausibly be extended another 2 bases at the 3′ end (Fig. 9A). In general, the CR1 repeats are located in analogous positions in Bongo, PegLeg, and Rey (Fig. 6 to 8). An attractive plausible role of these sequences is in the regulation of transcription, for the following reasons. First, in each instance the orientation of these asymmetric sequences corresponds with the direction of transcription of the genes immediately downstream. Second, they are predominantly intergenic, although in both CR1-2 and CR1-4 the termination codon of the upstream gene lies within the conserved sequence (Fig. 9A). Third, two of them (CR1-3 and CR1-5) are positioned immediately downstream of a putative terminator, where they are positioned potentially to initiate transcription of the downstream genes. Whether these sequences function directly in positioning transcription initiation or act as sites for positive or negative regulation of transcription is unclear, but we note that some of the sites are adjacent to plausible weak promoters; however, for two of the repeats (CR1-2 and CR1-4), there is little additional space to accommodate a promoter between the adjacent genes. Although the length of the repeat is similar to that for promoters of phage T7, there is no evidence for a phage-encoded RNA polymerase, and detailed transcriptome analysis is needed to elucidate their roles.

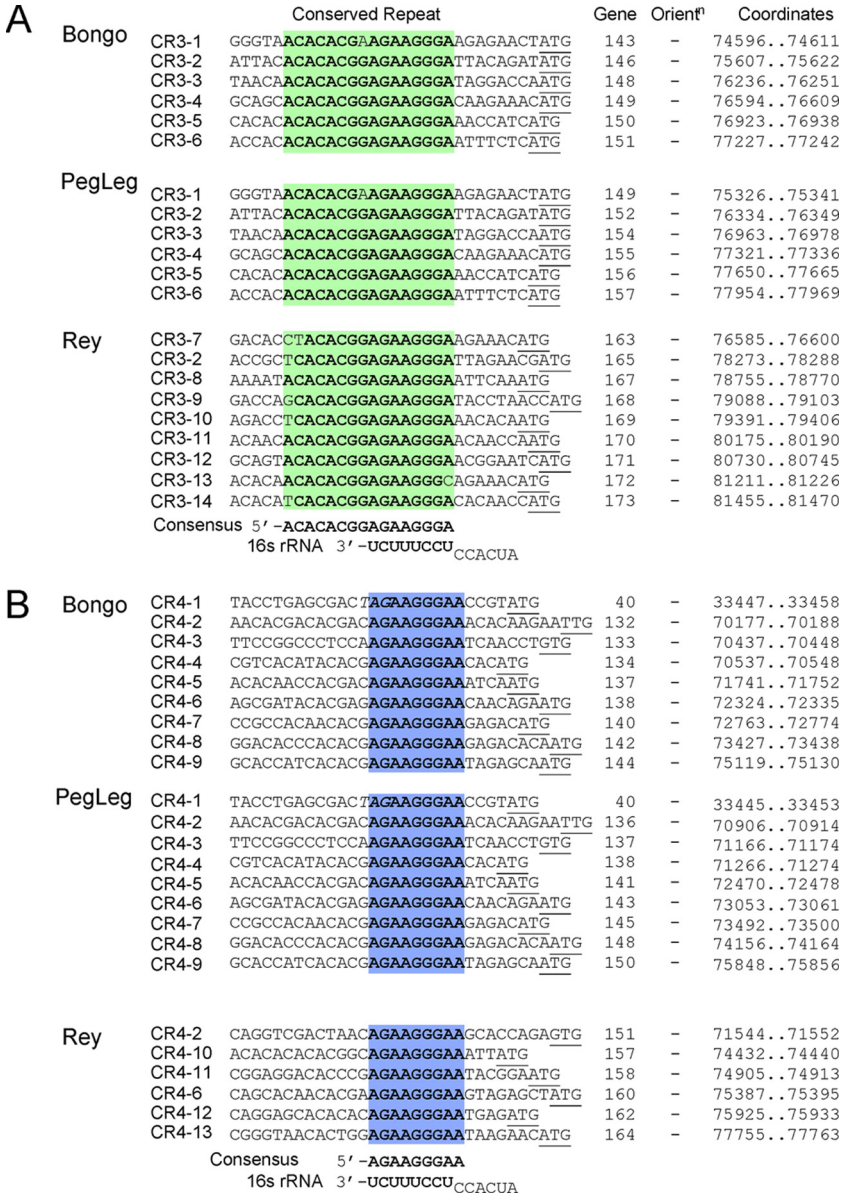


FIG 10 Conserved repeat sequences CR3 and CR4 in cluster M phages. Conserved repeated sequences CR3 (A) and CR4 (B) in Bongo, PegLeg, and Rey are shown, together with their flanking genes, the translation start codon of the downstream gene (underlined), orientation, and coordinates. CR3 and CR4 sequences are boxed in green and blue, respectively, along with their consensus sequences aligned with the 3′ end of the 16S rRNA of *M. smegmatis*.

A second set of conserved sequences contains a 12-bp sequence (5′-GTGATTACAGGA) that is closely related to the 3′ portion of CR1 (Fig. 9B). However, it differs from CR1, in that position 2 of CR2 is a nearly invariant T residue (with the exception of Rey CR2-4), and in CR1 this is a nearly invariant C base (with the exception of CR1-5 in Rey). In addition, the penultimate 3′ base of CR2 is an invariant G base, which corresponds to the variable position (purine) in CR1 (Fig. 9A). Like CR1, CR2 is predominantly intergenic (the only exception being CR2-9 in Rey, which is within the coding sequence of gene *85*). However, the correspondence with the orientation of transcription of the surrounding genes is inverted to that of CR1 and its surrounding genes. For example, in Bongo, CR2-4, CR2-5, and CR2-6 all contain the motif on the bottom strand, but the genes around CR2-4, CR2-5, and CR2-6 are all transcribed rightward; CR1-1, CR1-2, CR1-3, and CR1-4 are also among the rightward-transcribed genes, but the conserved motif is on the top strand (Fig. 6 and 9C). The role of CR2 and its relationship to CR1 are unclear. However, we note that the motif is not common in other mycobacteriophage genomes.

A third set of conserved repeats may play a role in regulation of translation initiation. In Bongo and PegLeg, there are six identical copies of the 16-bp asymmetric sequence 5′-ACACACGGAGAA GGGA (Fig. 10A), all located within the leftward-transcribed operon at the right end of the genome (Fig. 6 and 7). We refer to these as conserved repeat 3 (CR3). In Rey, there are three identical copies of the repeat, another five with a single-base mismatch, and one with a 2-base mismatch (there are no other sites with two

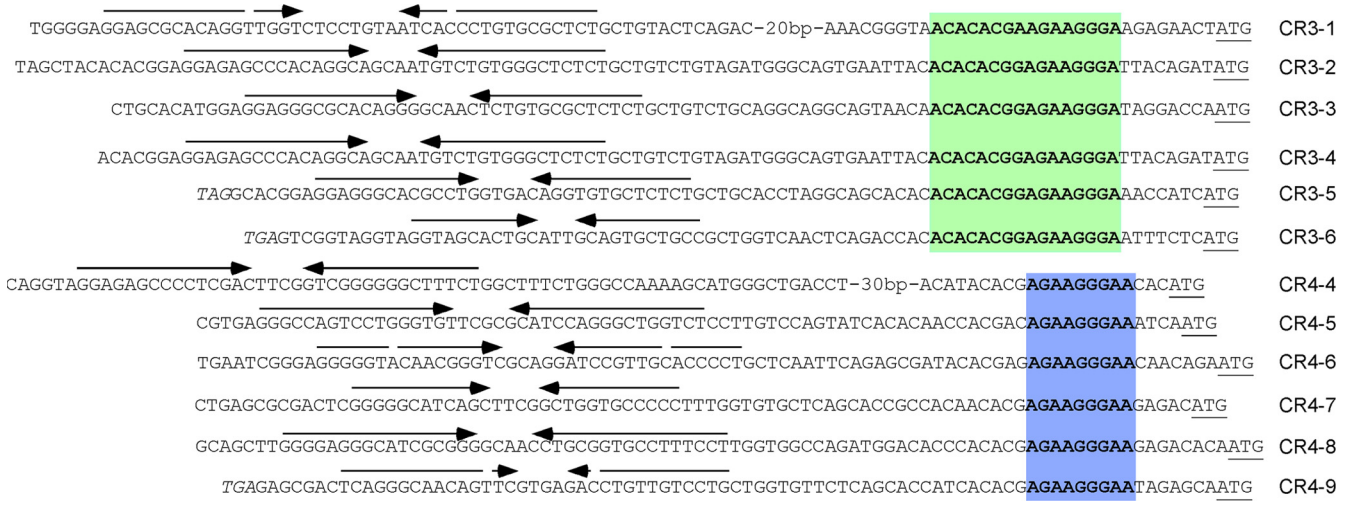


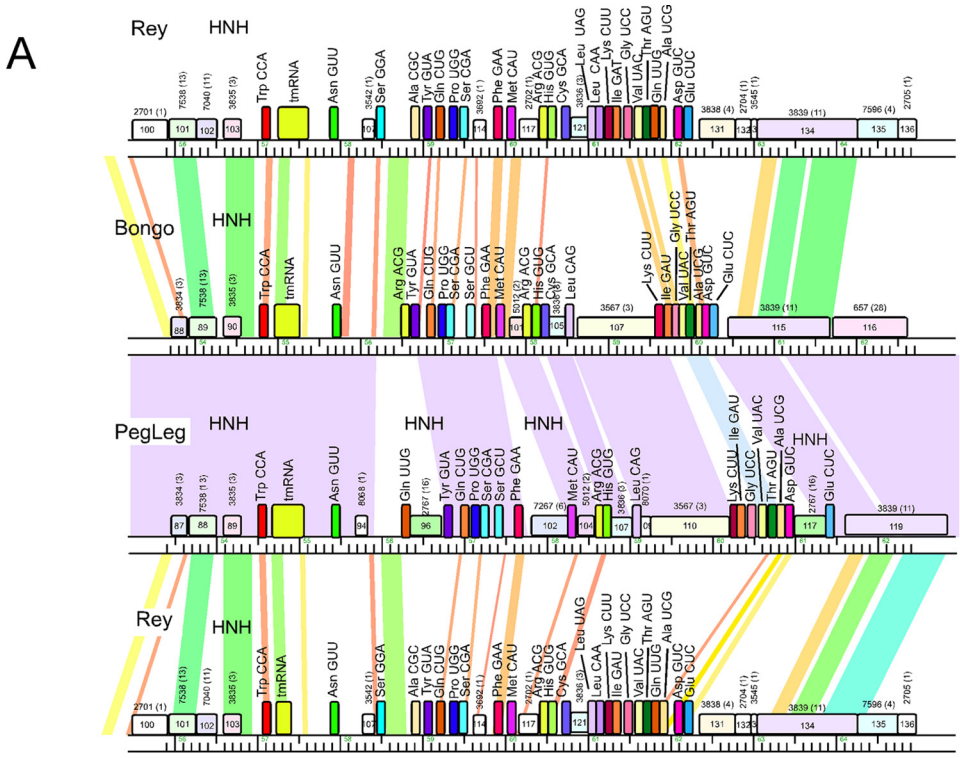
FIG 11 Conserved stem-loop sequences upstream of CR3 and CR4 in cluster M phages. CR3 sequences are boxed in green, and CR4 sequences are boxed in blue. Stem-forming sequences are shown by arrows under the nucleotide sequences.

mismatches or less in the genome; Fig. 10A); all are located in the leftward-transcribed operon at the right genome end (Fig. 9). These repeats are different but reminiscent of the start-associated sequences (SASs) reported previously in the cluster K mycobacteriophages (19) and have the following features. First, they are all located immediately upstream of putative translation initiation codons, spaced 6 to 8 bp from an ATG start (Fig. 10A). Second, they contain a 5′-AGAAGGGA motif that corresponds to a strong ribosome binding site and is perfectly complementary to the 8 bases at the extreme 3′ end of the 16S rRNA (Fig. 10A). Nonetheless, the conserved sequences extend beyond the 3′ end of the rRNA by an additional 8 bases, and it is unlikely that the sequence is simply conserved for ribosome binding through interactions with rRNA. The observation that these are constrained to a small region of the genome is consistent with them playing a regulatory role, perhaps through a phage-encoded function that modulates translation initiation at these genes. The finding that both cluster M and cluster K phages share these features—albeit with different conserved sequences—suggests that this may be a more widespread regulatory system than previously recognized.

A fourth set of conserved repeats (CR4) is also predicted to play a role in translation initiation (Fig. 10B). In Bongo and PegLeg, there are nine occurrences of the sequence 5′-AGAAGGGAA, and there are six in Rey. Like CR3, CR4 is positioned upstream of start codons (Fig. 10B), is 9 bp long, and is a partial subset of CR3, containing the rightmost 8 bp of CR3 but extended for 1 additional base (Fig. 10B). The 8 bases in common with CR3 are perfectly complementary to the 3′ end of the 16S rRNA, but the additional invariant A residue at the extreme right end is not (Fig. 10B). While the conservation of this 1 base does not provide compelling evidence for a regulatory system in addition to simply promoting ribosome binding for translation initiation, we note that CR4 is located primarily within the leftward-transcribed operon at the genome right end, with the one exception being upstream of Bongo gene *40*. We also note that CR4 is distinct from CR3, in that the spacing from the start codon is more variable (3 to 9 bp) and the start codon is not always ATG, as seen with CR3 (Fig. 10).

In the cluster K phages, a subset of the genes with start-associated sequences is accompanied by an extended start-associated sequence (ESAS) characterized by a conserved inverted repeat (19). Its role is unclear, and the lack of potential G·U base pairs precludes a clear indication of whether it represents a DNA recognition site for a regulatory protein or plays a role in RNA folding. We thus examined the cluster M phages for similar features upstream of CR3 and CR4. Although an ESAS as such was not observed, all six of the CR3 motifs in Bongo are associated with an upstream inverted repeat located within the intergenic space 20 to 40 bp upstream of CR3 (Fig. 11). Six of the CR4 repeats are associated with a similar inverted repeat, and the three that do not (CR4-1, CR4-2, and CR4-3) have only short intergenic upstream spaces. The inverted repeats have the potential to form 12- to 16-bp paired stems in RNA, accommodating G·U pairing; most form perfectly paired stems, and only three have any mispaired bases in the stem (Fig. 11). None of these repeats contains a run of thymines to their 3′ sides and thus are not predicted to function as transcriptional terminators. We propose that they may play a role in conjunction with the CR3 and CR4 motifs in regulation of the downstream genes, but the mechanism is unclear.

Clusters of tRNA genes in Bongo, Rey, and PegLeg. The cluster M phages have a set of 21 to 24 predicted tRNA genes together with a transfer-messenger RNA (tmRNA) gene located within a 5- to 7-kbp span and interspersed with small open reading frames (Fig. 12A; Tables 3 to 5). Most of these tRNAs are predicted with a reasonably high degree of confidence, identified by both the Aragorn and tRNAscan-SE programs with a Cove score of >20, although 4 to 6 are predicted with lower confidence because they either were not reported by Aragorn or have lower tRNAscan-SE Cove scores (Tables 3 to 5). Although these could be false-positive predictions, we have included them because phage tRNAs may have noncanonical structures or atypical modifications to escape host-mediated destruction (51). This is illustrated by comparison of the cluster M phage tRNAs with those of *M. smegmatis* mc^{2}155. Using tRNAscan-SE and including only those tRNAs with Cove scores above a threshold value of 20, the 47 predicted *M. smegmatis* tRNAs have a mean Cove score of 74.30. In contrast, the Bongo, PegLeg, and Rey tRNAs have mean Cove scores of 50.25, 47.91, and 45.70, respec-



B

Bongo tRNA	PegLeg	Rey
trp CCA	100	67
tmRNA	100	52
Asn GUU	100	<35
Tyr GUA	<35	37
Gln CUG	100	45
Pro UGG	100	<35
Ser CGA	100	<35
Ser GCU	100	x
Phe GAA	100	<35
Met CAU	<35	83
Arg ACG	100	<35
His GUG	97	<35
Cys GCA	x	<35
Leu CAG	96	x
Lys CUU	100	<35
Ile GAU	100	<35
Gly UCC	<35	92
Val UAC	<35	84
Thr AGU	100	<35
Ala CGA	100	97
Asp GUC	88 (96)	88 (96)
Glu CUC	<35	<35

C

tRNA	Bongo	PegLeg	Rey
Ser GGA			O
Gln UUG		O	
Ala CGC			O
Arg ACG	O		
Leu UAG			O
Leu CAA			O
Gln UUG			O

D

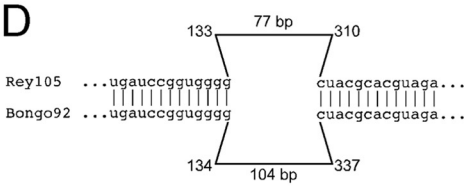


FIG 12 Organizations of tRNA genes in cluster M mycobacteriophages. (A) tRNA cluster from each of the cluster M phages shown in a pairwise comparison to each of the others, as drawn by Phamerator. tRNAs are colored by isotype. (B) Nucleotide sequence similarity of PegLeg and Rey tRNAs relative to Bongo tRNAs. tRNA isotypes are listed in the far left column. Boxes with an X indicate that a genome does not contain that tRNA isotype. Two numbers are shown for Asp (GUC); the number in parentheses is the similarity of the PegLeg and Rey tRNAs to each other rather than to the tRNA of Bongo. This tRNA was the only one in which the PegLeg-Rey pair was more similar to each other than either of them was to the Bongo isotype; thus, the remaining isotypes are shown in comparison only to those in Bongo. (C) Single tRNA isotypes that appear only in one location in one cluster M genome and not in the corresponding locations in the others. (D) Alignment of the tmRNAs in Rey and Bongo. The tmRNAs are identical at the 5′ and 3′ ends, with a large dissimilar segment in the center of the genes.

tively. The lower overall Cove scores for phage tRNAs may be a reflection of the lower selective pressure on the phages compared to that on the hosts, a phenomenon that has been noted with respect to conserved residues of phage/host protein homologs. It remains to be determined if all the phage-encoded tRNAs are functional, but atypical tRNA structures appear to be common for phage-encoded tRNAs. Some of the tRNAs have noncanonical base positions or mismatches in stem pairings (Tables 3 to 5), but a comparison of the three genomes suggests that many of these are indeed functional (Fig. 12). For example, a $tRNA^{Tyr}$(GUA) is con-

TABLE 3 Predicted Bongo tRNAs

Prediction confidence and isotype[a]	Anticodon	Coordinates	Aragorn[b]	Cove score[c]	Comments[d]
Higher-confidence predictions					
Trp	CCA	54782–54853	Y	45.39	U2:G72, U31:G39
Asn	GUU	55702–55773	Y	45.49	C9, G11, C24
Gln	CUG	56854–56926	Y	45.18	C9, A32
Pro	UGG	56930–57003	Y	41.05	A8, G21, G27:G43
Phe	GAA	57490–57565	Y	67.61	U5:U68
Met	CAU	57653–57724	Y	55.32	
Arg	ACG	58004–58078	Y	40.08	C8, G14, G21, C27:A43
His	GUG	58081–58152	Y	44.41	T30:C40
Leu	CAG	58492–58566	Y	57.37	
Lys	CUU	59586–59659	Y	54.24	
Gly	UCC	59782–59857	Y	39.55	
Val	UAC	59862–59934	Y	64.11	G48
Thr	AGU	59938–60010	Y	69.47	U6:U67
Asp	GUC	60125–60200	Y	54.17	C9, C19
Glu	CUC	60202–60273	Y	55.20	
Lower-confidence predictions					
Arg	ACG	56531–56604	Y	8.03	U6:U67, C10, 49A:65G, 30G:40A, C33
Tyr	GUA	56609–56686	Y	11.78	C8, A11:C24
Ser	CGA	57044–57127	N	4.09	C10, G11, C10:C25, G11:G24, C22, C31:A39, G49:A65
Ser	GCU	57299–57380	Y	N	U4:U69, A8, C9, C10, A11, U24, G25, A54, U58
Cys	GCA	58158–58230	Y	16.54	G54, G55, G58
Ile	GAU	59660–59734	N	25.40	U1:U72, U19, G54, G55, A60
Ala	UGC	60014–60087	N	9.25	A8, U9, U11:G24, U12:G23, C14, C15

[a] tRNAs listed with higher confidence are predicted by the Aragorn and by the tRNAscan-SE programs with a score of >20; lower-confidence tRNAs are predicted either by the Aragorn program or by the tRNAscan-SE program with a score of >4.
[b] tRNAs were predicted using the Aragorn (v1.2.36) program. Y, yes; N, no.
[c] tRNAs were predicted using the tRNAscan-SE (v1.21) program, and Cove scores were identified.
[d] Departures from well-conserved base positions are noted, using standard numbering (the anticodon is at positions 34 to 36). The well-conserved positions are U8, R9, R10, Y11, A14, R15, G18, G19, A21, R24, Y25, Y32, U33, R37, Y48, R52, G53, U54, U55, R57, A58, Y60, C61, and Y62. Noncanonical base pairs are shown as N:N.

served in all three cluster M genomes at syntenically related positions, even though these tRNAs are quite divergent at the nucleotide sequence level (Fig. 12B). Each has different noncanonical bases or base pairings, but the conservation of isotype and gene order suggests that they are nonetheless functional. Similar patterns are seen with other predicted tRNA genes (Fig. 12B; Tables 3 to 5).

The tRNA locus is the region of the greatest divergence between Bongo and PegLeg, and both are only weakly related to Rey in this region (Fig. 12). PegLeg differs largely due to four HNH genes, three of which are absent from Bongo and Rey. Interestingly, there are five tRNAs [tRNA^{Tyr}(GUA), tRNA^{Met}(CAU), tRNA^{Gly}(TCC), tRNA^{Val}(UAC), and tRNA^{Glu}(CUC)] where the PegLeg sequences have diverged greatly from those in Bongo (<35% identity)—although the isotypes are retained—and have likely been acquired by horizontal genetic exchange; in addition, PegLeg has a tRNA^{Gln}(UUG) that Bongo lacks. Four of these [tRNA^{Tyr}(GUA), tRNA^{Met}(CAU), tRNA^{Glu}(CUC), and tRNA^{Gln}(UUG)] are immediately adjacent to an HNH gene, and the other two are nearby (Fig. 12). These HNH endonucleases are thus directly implicated in generating diversity within the tRNA repertoire, perhaps facilitating rapid adaptation to growth in new bacterial hosts that have poorly matched host tRNA abundances with the phage requirements. This is reminiscent of the role of SegB in tRNA inheritance in phage T4 (52). We note than in spite of the low overall sequence similarity between Rey and the subcluster M1 phages in this region, 17 of the tRNA isotypes are present and in syntenically conserved positions. Although Rey lacks the M1 phages' tRNA^{Ser}(CGA), tRNA^{Ser}(GCU), and tRNA^{Leu}(CAG), it has acquired tRNA^{Ser}(GGA), tRNA^{Ala}(CGC), tRNA^{Leu}(UAG), and tRNA^{Leu}(CAA) (Fig. 12C).

Interestingly, each genome encodes tRNAs for all 20 isotypes, with the exception of PegLeg, for which no tRNA^{Cys} is predicted (Fig. 13). However, this distribution is somewhat odd, as there appears to be less than a complete coding set of tRNAs. Even if wobble paring, superwobble pairing (53), and both canonical and atypical tRNA modifications facilitate the reading of each two-box, three-box, and four-box codon set, at least three additional tRNAs would be required per genome to read the six-box codon sets (Leu, Arg; Ser; Fig. 13). Both Bongo and PegLeg encode two serine isotype tRNAs [tRNA^{Ser}(CGA) and tRNA^{Ser}(GCU)], but they do not encode tRNA for the Leu and Arg codons AGR and UUR, respectively (Fig. 13). Rey differs in that it lacks the tRNA^{Ser}(GCU), but it also encodes a tRNA^{Leu}(CAA) (Fig. 13). We also note that even decoding of all two-box, three-box, and four-box codon sets would require unusual tRNA modifications. For example, the Bongo and PegLeg tRNA^{Gly}(UCC) and tRNA^{Ala}(UGC) would not typically be expected to read all four glycine and alanine codons, respectively, and there is no tRNA for decoding the isoleucine AUA codon [Rey has a tRNA^{Ala}(CGC)].

TABLE 4 Predicted PegLeg tRNAs

Isotype[a]	Anticodon	Coordinates	Aragorn[b]	Cove score[c]	Comments[d]
Higher-confidence predictions					
Trp	CCA	54522–54595	Y	45.39	U2:G72, U31:G39
Asn	GUU	55442–55513	Y	45.49	C9, G11, C24
Gln	UUG	56269–56340	Y	31.65	A5:G68, A32, A33
Tyr	GUA	56725–56807	Y	30.23	C(Ins 7–8), C9, C10, U15,G25
Gln	CUG	56976–57048	Y	45.18	C9, A32
Pro	UGG	57051–57125	Y	41.05	A8, A21, G27:G43
Phe	GAA	57612–57687	Y	67.61	U5:U68
Met	CAU	58249–58324	Y	67.76	A11, U21, U24
Arg	ACG	58604–58678	Y	40.08	C8, G14, C21, C27:A43
His	GUG	58681–58753	Y	50.45	
Leu	CAG	59015–59089	Y	56.01	G11, C24
Lys	CUU	60212–60285	Y	54.24	
Gly	UCC	60444–60520	Y	48.73	U14
Val	UAC	60573–60644	Y	39.94	C9, G32
Thr	AGU	60706–60778	Y	69.47	U6:U67
Asp	GUC	60894–60967	Y	41.67	C9, A19
Glu	CUC	61355–61429	Y	62.10	A11, U24
Lower-confidence predictions					
Ser	CGA	57166–57249	N	4.09	C10, G11, C10:C25, G11:G24, C22, C31:A39, G49:A65
Ser	GCU	57421–57502	Y	N	U4:U69, A8, C9, C10, A11, U24, G25, A54, U58
Ile	GAU	60286–60360	N	25.40	U1:U72, U19, G54, G55, A60
Ala	UGC	60782–60855	N	9.25	A8, U9, U11:G24, U12:G23, C14, C15

[a] tRNAs listed with higher confidence are predicted by the Aragorn and by the tRNAscan-SE programs with a score of >20; lower-confidence tRNAs are predicted either by the Aragorn program or by the tRNAscan-SE program with a score of >4.
[b] tRNAs were predicted using the Aragorn (v1.2.36) program. Y, yes; N, no.
[c] tRNAs were predicted using the tRNAscan-SE (v1.21) program, and cove scores were identified.
[d] Departures from well-conserved base positions are noted, using standard numbering (the anticodon is at positions 34 to 36). The well-conserved positions are U8, R9, R10, Y11, A14, R15, G18, G19, A21, R24, Y25, Y32, U33, R37, Y48, R52, G53, U54, U55, R57, A58, Y60, C61, and Y62. Noncanonical base pairs are shown as N:N.

Comparison of the normalized synonymous codon usage frequencies of the cluster M phages and *M. smegmatis* shows similar codon usage preferences (Fig. 14). However, there is a notable difference in codon usage between the rightward-transcribed operon containing virion structure and assembly genes (e.g., Bongo genes *12* to *39*; Fig. 13) and the rest of the genome (Fig. 13). For example, although this operon contains 42% of the total codons, in Bongo (Fig. 13) and PegLeg (Fig. 13), it is almost completely depleted of UUA (Leu), AUA (Ile), and the AGR (Arg) codons (in Bongo, the only AUA codon is found in gene *12*, the first gene transcribed in the forward direction and of unknown function, and the two AGG codons are in an early and possibly untranslated part of gene *22* and in an untranslated part of gene *25*, a reading frame expressed as the downstream part of the programmed translational frameshift). Thus, late gene expression may be almost entirely dependent on the phage-encoded tRNAs, and codons for which phage-encoded tRNAs are lacking are absent. In Rey, the codon distribution is somewhat different (Fig. 13), in that there are 14 AGG codons in the late operon, and either it presumably uses a host tRNA for these or there may be an unidentified noncanonical Rey-encoded tRNA for this function.

The tmRNAs encoded by Bongo and PegLeg are identical but differ substantially from the Rey tmRNA. The 5′ 133 nucleotides share 94% nucleotide sequence identity, and the 3′ 66 nucleotides are identical, but the central ~200-bp segments are virtually unrelated (Fig. 12D). The degradation tag is encoded within the common 5′ region, and the sequence departures give rise to slightly different tags (AAFVDADYAVAA in Bongo and PegLeg and AAFVAADYAVGA in Rey), each of which differs from the host tmRNA-encoded tag (ADSNQRDYALAA). We note that all three cluster M phages also encode a Clp-like protease (e.g., Bongo gp58), which could be specifically involved in degrading stalled tmRNA-tagged proteins and releasing ribosomes for translation of other mRNAs.

DISCUSSION

As the number of completely sequenced mycobacteriophages increases, new genomes that are distinct at the nucleotide sequence level from previously described phages are isolated less frequently. The cluster M phages thus represent an interesting new group displaying a variety of novel features. We have suggested previously that the broad span of mycobacteriophage diversity and their broad range of GC contents reflects an evolutionary history in which phages migrate rapidly across a landscape of closely related but diverse bacterial hosts at relatively high frequencies (20). The cluster M phages may thus have accessed *M. smegmatis* mc^{2}155 as a host in their recent evolutionary past but took a distinct evolutionary path, accessing a spectrum of hosts that differs from that of mycobacteriophage clusters. The various features of the cluster M phages thus likely reflect not just their current adaptations to propagate in *M. smegmatis* mc^{2}155 but also their requirements for prior hosts in their evolution. The genomic differences between the subcluster M1

TABLE 5 Predicted Rey tRNAs

Isotype[a]	Anticodon	Coordinates	Aragorn[b]	Cove score[c]	Comments[d]
Higher-confidence predictions					
Trp	CCA	57088–57160	Y	36.86	G14
Asn	GUU	57913–57986	Y	55.69	U6:U67
Ser	GGA	58475–58548	Y	50.30	A1:A72, A2:A71, U19, G55
Ala	CGC	58843–58915	Y	31.32	A5:C68, C10, A11:C24, G12:G23, C14, U17, Δ26
Gln	CUG	59197–59268	Y	33.32	U7:C66, A32
Pro	UGG	59359–59434	Y	41.36	A8, Δ9, G11, G10:G25, G21, Δ26, A27,G43
Phe	GAA	59833–59908	Y	56.09	U4:U69, U5:U68
Met	CAU	60048–60120	Y	47.54	
Arg	ACG	60409–60487	Y	50.39	U9, G21
His	GUG	60490–60564	Y	40.18	A54, U58, A60
Cys	GCA	60672–60744	Y	38.12	A28:A42
Leu	UAG	60998–61072	Y	42.05	U50:U64
Leu	CAA	61076–61151	Y	51.68	G11, A28:A42
Lys	CUU	61189–61260	Y	54.65	U21, U30:U40
Ile	GAU	61261–61336	Y	47.92	U1:U72, U6:U67
Gly	UCC	61425–61500	Y	40.71	C9
Val	UAC	61563–61635	Y	61.45	G48
Thr	AGU	61696–61773	Y	58.76	U9, U15, U21
Asp	GUC	62051–62124	Y	44.73	C9
Glu	CUC	62184–62255	Y	30.93	C27:U43, A54, A58, A60
Lower-confidence predictions					
Tyr	GUA	59005–59082	Y	8.14	Non-canonical D-loop, G49:A65
Ser	CGA	59472–59556	N	5.35	C8:Δ9, C12:U28, G25, C31:A39, G32, A48, G49:A65, G51:T62
Gln	UUG	61775–61848	N	19.77	C9, G10:U25, A54
Ala	UGC	61885–61928	N	12.09	C7:U66, A8, U9, U12: G23C14, U15, G19, non-canonical anticodon stem, G48

[a] tRNAs listed with higher confidence are predicted by the Aragorn and by the tRNAscan-SE programs with a score of >20; lower-confidence tRNAs are predicted either by the Aragorn program or by the tRNAscan-SE program with a score of >4.
[b] tRNAs were predicted using the Aragorn (v1.2.36) program. Y, yes; N, no.
[c] tRNAs were predicted using the tRNAscan-SE (v1.21) program, and cove scores were identified.
[d] Departures from well-conserved base positions are noted, using standard numbering (the anticodon is at positions 34 to 36). The well-conserved positions are U8, R9, R10, Y11, A14, R15, G18, G19, A21, R24, Y25, Y32, U33, R37, Y48, R52, G53, U54, U55, R57, A58, Y60, C61, and Y62. Noncanonical base pairs are shown as N:N.

and M2 phages perhaps reflect differences in the particular hosts accessed within these evolutionary pathways. Likewise, although phage Wildcat does not meet the threshold criteria for inclusion in cluster M, it not only has a similar genomic architecture but also shares most of the other cluster M features, including a similar tRNA repertoire.

The four sets of conserved repeated sequences suggest intriguing possibilities for regulatory systems. In general, they are all located in small intergenic regions and as such are reminiscent of the stoperators in the cluster A phages that are sites for repressor binding and silencing of transcription (14, 17). The CR1 and CR2 repeats could be involved in transcription initiation or regulation, and one or both could function as operator or stoperator sites, although neither the promoters that could be regulated nor the identity of the repressor genes is readily apparent. The CR3 and CR4 repeats are more likely to be involved in translational regulation, as they are closely linked to predicted translational start sites and are reminiscent of the SASs of the cluster K phages (19). As with the cluster K SAS sites, the CR3 and CR4 repeats are devoid from the late structural genes and are primarily restricted to the leftward-transcribed operon at the right end of the cluster M genomes. While these could be involved in promoting translation initiation, it is also plausible that they mediate downregulation of these genes during late gene expression so as to optimize the availability of ribosomes for structural gene synthesis.

Analysis of the virion structural proteins by mass spectrometry revealed the presence of three proteins that we did not expect to find, an Erf-like protein, a Clp-like protease, and SSB. Although we cannot exclude the possibility that these are contaminants arising from their abundance, it seems more likely that they are located within the capsid. It is plausible that Erf and SSB play roles in recircularization of the genome during infection, although the genome contains cohesive termini similar to those of many other phages, and a requirement for these functions is atypical. The protease could plausibly remove host proteins that may bind to the genome ends and inhibit circularization or perhaps modulate the activities of the Erf-like and SSB proteins.

The tRNAs encoded by other mycobacteriophages—such as L5, D29, and Bxz1—have been interpreted in the context of the codon usage of the phage relative to that of the host (54–57). However, given the unusual set of tRNAs in the cluster M phages, we favor an alternative explanation. It has been shown previously that bacterial hosts evolve mechanisms for phage resistance that involve cleavage of essential tRNAs and that phages can counter this by encoding their own tRNAs that resist cleavage (through

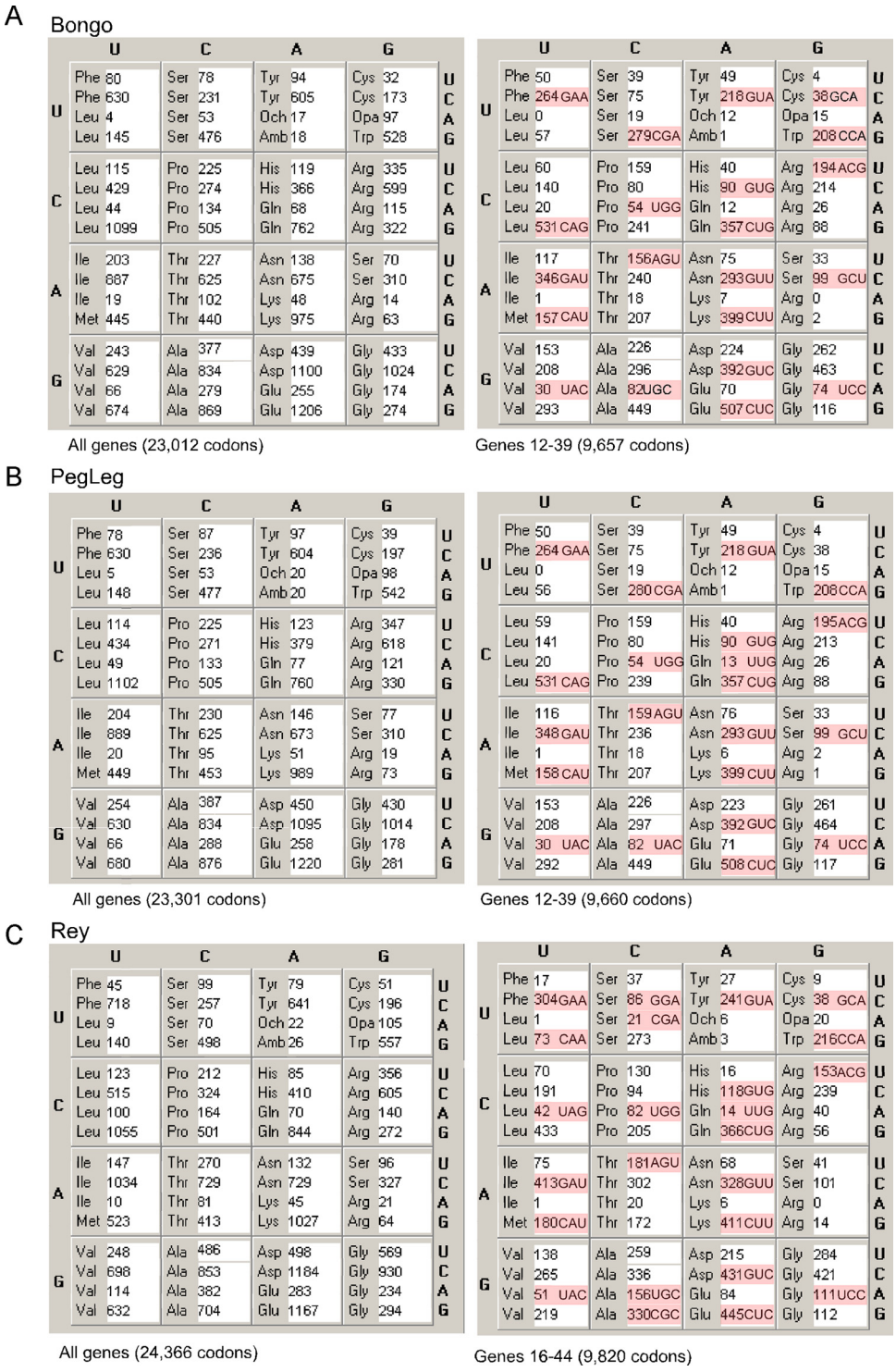


FIG 13 Codon usage in mycobacteriophages Bongo, PegLeg, Rey and *M. smegmatis* mc^{2}155. The left box shows the codon counts for all predicted protein-coding genes in Bongo (A), and the right box shows those in the putative structural gene operon containing genes *12* to *39*. The codons for which there is a corresponding tRNA are shown in red, with the anticodon indicated. The total numbers of codons are shown below. Similar analyses are shown for PegLeg (B) and Rey (C), with the putative structural operon in Rey containing genes *16* to *44*.

unusual bases, base pairings, or modifications). As all of the cluster M phages encode an RtcB RNA ligase implicated in tRNA repair (48), a similar scenario is plausible. Because these phages encode a large set of tRNAs, it is possible that all or most of the host tRNAs are destroyed and that the phage-encoded tRNAs are the predominant ones used during infection. Destruction of host tRNAs is not expected to occur in *M. smegmatis*, as most other phages do not encode large tRNA sets but could have occurred in

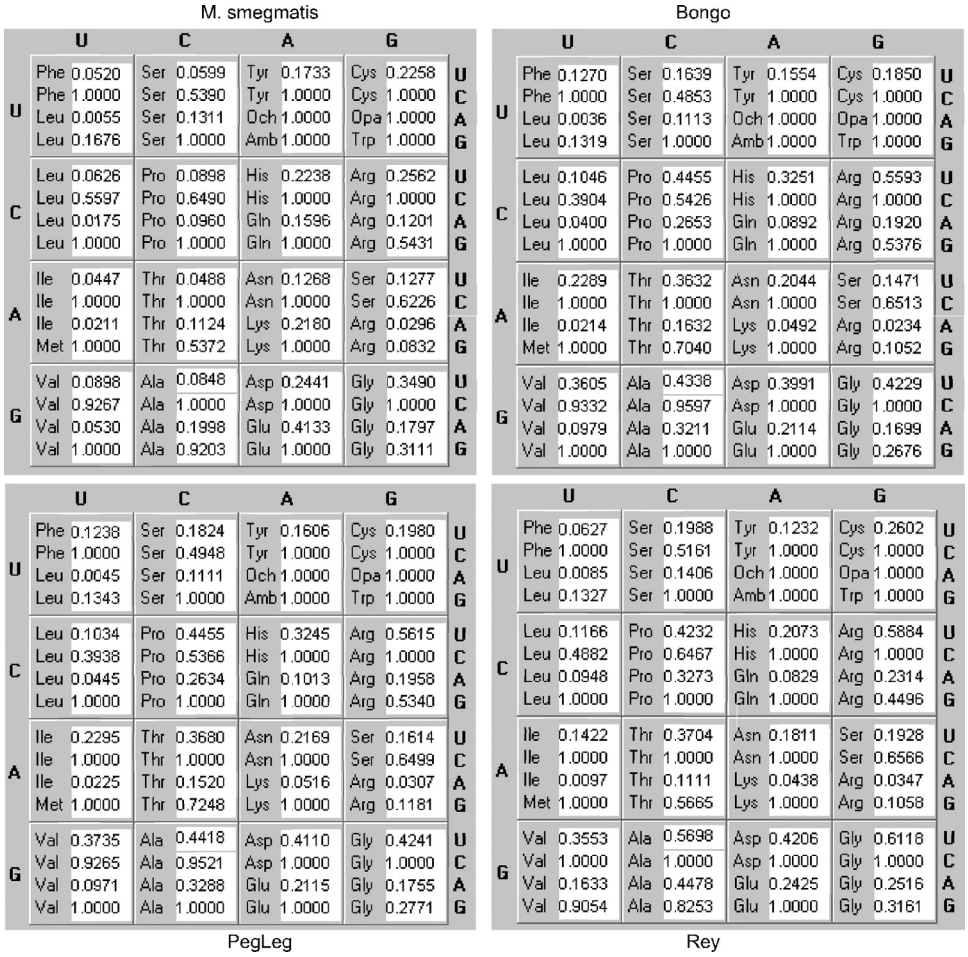


FIG 14 Normalized synonymous codon usage in *Mycobacterium smegmatis* mc^{2}155, Bongo, PegLeg, and Rey. For each genome, the codons of all protein-encoding genes were totaled and normalized relative to the most prevalent codon for each amino acid for that genome. The frequency of the most common codon is reported to be 1.0000.

a different host in the recent evolutionary past of the cluster M phages, providing a mechanism for resistance through abortive infection.

ACKNOWLEDGMENTS

We thank Tracy Le, Jan Michael Taguiam, Jonathan Tran, and Junhao (Jay) Zhang for the isolation of PegLeg and Josephine Nguyen and Sean Chang for the initial genome characterization of PegLeg. Special thanks go to the UCLA instructional team of ERS, Krisanavane Reddi, William Villella, Todd C. Lorenz, and Janahan Vijanderan, for overseeing all aspects of the discovery and genome analysis process.

This research project was supported in part by the Howard Hughes Medical Institute SEA-PHAGES program and Howard Hughes Medical Institute Precollege and Undergraduate Science Education Program grants to the University of California, Los Angeles (award no. 52006944), and Washington University in St. Louis, MO (award no. 52006961).

REFERENCES

1. **Hatfull GF, Hendrix RW.** 2011. Bacteriophages and their genomes. Curr. Opin. Virol. **1:**298–303. http://dx.doi.org/10.1016/j.coviro.2011.06.009.
2. **Hendrix RW.** 2009. Jumbo bacteriophages. Curr. Top. Microbiol. Immunol. **328:**229–240. http://dx.doi.org/10.1007/978-3-540-68618-7_7.
3. **Hatfull GF.** 2010. Mycobacteriophages: genes and genomes. Annu. Rev. Microbiol. **64:**331–356. http://dx.doi.org/10.1146/annurev.micro.112408.134233.
4. **Hatfull GF.** 2012. The secret lives of mycobacteriophages. Adv. Virus Res. **82:**179–288. http://dx.doi.org/10.1016/B978-0-12-394621-8.00015-7.
5. **Pope WH, Jacobs-Sera D, Best AA, Broussard GW, Connerly PL, Dedrick RM, Kremer TA, Offner S, Ogiefo AH, Pizzorno MC, Rockenbach K, Russell DA, Stowe EL, Stukey J, Thibault SA, Conway JF, Hendrix RW, Hatfull GF.** 2013. Cluster J mycobacteriophages: intron splicing in capsid and tail genes. PLoS One **8:**e69273. http://dx.doi.org/10.1371/journal.pone.0069273.
6. **Hatfull GF.** 2012. Complete genome sequences of 138 mycobacteriophages. J. Virol. **86:**2382–2384. http://dx.doi.org/10.1128/JVI.06870-11.
7. **Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM, Hryckowian AJ, Kelchner VA, Namburi S, Pajcini KV, Popovich MG, Schleicher DT, Simanek BZ, Smith AL, Zdanowicz GM, Kumar V, Peebles CL, Jacobs WR, Jr, Lawrence JG, Hendrix RW.** 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. **2:**e92. http://dx.doi.org/10.1371/journal.pgen.0020092.
8. **Hatfull GF.** 1994. Mycobacteriophage L5: a toolbox for tuberculosis. ASM News **60:**255–260.
9. **Hatfull GF.** 2004. Mycobacteriophages and tuberculosis, p 203–218. *In* Eisenach K, Cole ST, Jacobs WR, Jr, McMurray D (ed), Tuberculosis. ASM Press, Washington, DC.
10. **Banaiee N, Bobadilla-Del-Valle M, Bardarov S, Jr, Riska PF, Small PM, Ponce-De-Leon A, Jacobs WR, Jr, Hatfull GF, Sifuentes-Osornio J.** 2001. Luciferase reporter mycobacteriophages for detection, identifica-

tion, and antibiotic susceptibility testing of Mycobacterium tuberculosis in Mexico. J. Clin. Microbiol. **39:**3883–3888. http://dx.doi.org/10.1128/JCM.39.11.3883-3888.2001.

11. **Bonnet J, Subsoontorn P, Endy D.** 2012. Rewritable digital data storage in live cells via engineered control of recombination directionality. Proc. Natl. Acad. Sci. U. S. A. **109:**8884–8889. http://dx.doi.org/10.1073/pnas.1202344109.
12. **Nkrumah LJ, Muhle RA, Moura PA, Ghosh P, Hatfull GF, Jacobs WR, Jr, Fidock DA.** 2006. Efficient site-specific integration in Plasmodium falciparum chromosomes mediated by mycobacteriophage Bxb1 integrase. Nat. Methods **3:**615–621. http://dx.doi.org/10.1038/nmeth904.
13. **Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR, Hendrix RW, Hatfull GF.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171–182. http://dx.doi.org/10.1016/S0092-8674(03)00233-2.
14. **Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, Alexander LM, Alfano MB, Alford ST, Amy NE, Anderson MD, Anderson AG, Ang AAS, Ares M, Jr, Barber AJ, Barker LP, Barrett JM, Barshop WD, Bauerle CM, Bayles IM, Belfield KL, Best AA, Borjon A, Jr, Bowman CA, Boyer CA, Bradley KW, Bradley VA, Broadway LN, Budwal K, Busby KN, Campbell IW, Campbell AM, Carey A, Caruso SM, Chew RD, Cockburn CL, Cohen LB, Corajod JM, Cresawn SG, Davis KR, Deng L, Denver DR, Dixon BR, Ekram S, Elgin SCR, Engelsen AE, English BEV, Erb ML, Estrada C, Filliger LZ, et al.** 2011. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PLoS One **6:**e16329. http://dx.doi.org/10.1371/journal.pone.0016329.
15. **Sampson T, Broussard GW, Marinelli LJ, Jacobs-Sera D, Ray M, Ko CC, Russell D, Hendrix RW, Hatfull GF.** 2009. Mycobacteriophages BPs, Angel and Halo: comparative genomics reveals a novel class of ultra-small mobile genetic elements. Microbiology **155:**2962–2977. http://dx.doi.org/10.1099/mic.0.030486-0.
16. **Mageeney C, Pope WH, Harrison M, Moran D, Cross T, Jacobs-Sera D, Hendrix RW, Dunbar D, Hatfull GF.** 2012. Mycobacteriophage Marvin: a new singleton phage with an unusual genome organization. J. Virol. **86:**4762–4775. http://dx.doi.org/10.1128/JVI.00075-12.
17. **Brown KL, Sarkis GJ, Wadsworth C, Hatfull GF.** 1997. Transcriptional silencing by the mycobacteriophage L5 repressor. EMBO J. **16:**5914–5921. http://dx.doi.org/10.1093/emboj/16.19.5914.
18. **Broussard GW, Oldfield LM, Villanueva VM, Lunt BL, Shine EE, Hatfull GF.** 2013. Integration-dependent bacteriophage immunity provides insights into the evolution of genetic switches. Mol. Cell **49:**237–248. http://dx.doi.org/10.1016/j.molcel.2012.11.012.
19. **Pope WH, Ferreira CM, Jacobs-Sera D, Benjamin RC, Davis AJ, DeJong RJ, Elgin SCR, 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.** 2011. Cluster K mycobacteriophages: insights into the evolutionary origins of mycobacteriophage TM4. PLoS One **6:**e26750. http://dx.doi.org/10.1371/journal.pone.0026750.
20. **Jacobs-Sera D, Marinelli LJ, Bowman C, Broussard GW, Guerrero Bustamante C, Boyle MM, Petrova ZO, Dedrick RM, Pope WH, Science Education Alliance Phage Hunters Advancing Genomics Evolutionary Science (SEA-PHAGES) Program, Modlin RL, Hendrix RW, Hatfull GF.** 2012. On the nature of mycobacteriophage diversity and host preference. Virology **434:**187–201. http://dx.doi.org/10.1016/j.virol.2012.09.026.
21. **Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF.** 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc. Natl. Acad. Sci. U. S. A. **96:**2192–2197. http://dx.doi.org/10.1073/pnas.96.5.2192.
22. **Gordon D, Abajian C, Green P.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202. http://dx.doi.org/10.1101/gr.8.3.195.
23. **Delcher AL, Harmon D, Kasif S, White O, Salzberg SL.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641. http://dx.doi.org/10.1093/nar/27.23.4636.
24. **Borodovsky M, Lomsadze A.** 2011. Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. Curr. Protocols Bioinformatics Chapter 4:Unit 4.5.1-17. http://dx.doi.org/10.1002/0471250953.bi0405s35.
25. **Soding J, Biegert A, Lupas AN.** 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. **33:**W244–W248. http://dx.doi.org/10.1093/nar/gki408.
26. **Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF.** 2011. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics **12:**395. http://dx.doi.org/10.1186/1471-2105-12-395.
27. **Krumsiek J, Arnold R, Rattei T.** 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics **23:**1026–1028. http://dx.doi.org/10.1093/bioinformatics/btm039.
28. **Huson DH, Bryant D.** 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. **23:**254–267. http://dx.doi.org/10.1093/molbev/msj030.
29. **Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS.** 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. **37:**W202–W208. http://dx.doi.org/10.1093/nar/gkp335.
30. **Laslett D, Canback B.** 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. **32:**11–16. http://dx.doi.org/10.1093/nar/gkh152.
31. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
32. **Schattner P, Brooks AN, Lowe TM.** 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. **33:**W686–W689. http://dx.doi.org/10.1093/nar/gki366.
33. **Smith MC, Burns RN, Wilson SE, Gregory MA.** 1999. The complete genome sequence of the Streptomyces temperate phage phiC31: evolutionary relationships to other viruses. Nucleic Acids Res. **27:**2145–2155. http://dx.doi.org/10.1093/nar/27.10.2145.
34. **Smith MC, Hendrix RW, Dedrick R, Mitchell K, Ko CC, Russell D, Bell E, Gregory M, Bibb MJ, Pethick F, Jacobs-Sera D, Herron P, Buttner MJ, Hatfull GF.** 2013. Evolutionary relationships within actinophages and a putative adaptation for growth in Streptomyces spp. J. Bacteriol. **195:**4924–4935. http://dx.doi.org/10.1128/JB.00618-13.
35. **Mediavilla J, Jain S, Kriakov J, Ford ME, Duda RL, Jacobs WR, Jr, Hendrix RW, Hatfull GF.** 2000. Genome organization and characterization of mycobacteriophage Bxb1. Mol. Microbiol. **38:**955–970. http://dx.doi.org/10.1046/j.1365-2958.2000.02183.x.
36. **Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS, Salamon P, Youle M, Rohwer F.** 2013. Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc. Natl. Acad. Sci. U. S. A. **110:**10771–10776. http://dx.doi.org/10.1073/pnas.1305923110.
37. **Fraser JS, Yu Z, Maxwell KL, Davidson AR.** 2006. Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. **359:**496–507. http://dx.doi.org/10.1016/j.jmb.2006.03.043.
38. **Catalao MJ, Gil F, Moniz-Pereira J, Pimentel M.** 2010. The mycobacteriophage Ms6 encodes a chaperone-like protein involved in the endolysin delivery to the peptidoglycan. Mol. Microbiol. **77:**672–686. http://dx.doi.org/10.1111/j.1365-2958.2010.07239.x.
39. **Payne KM, Hatfull GF.** 2012. Mycobacteriophage endolysins: diverse and modular enzymes with multiple catalytic activities. PLoS One **7:**e34052. http://dx.doi.org/10.1371/journal.pone.0034052.
40. **Hendrix RW, Roberts JW, Stahl FW, Weisberg RA.** 1983. Lambda II. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
41. **Dedrick RM, Marinelli LJ, Newton GL, Pogliano K, Pogliano J, Hatfull GF.** 2013. Functional requirements for bacteriophage growth: gene essentiality and expression in mycobacteriophage Giles. Mol. Microbiol. **88:**577–589. http://dx.doi.org/10.1111/mmi.12210.
42. **Donnelly-Wu MK, Jacobs WR, Jr, Hatfull GF.** 1993. Superinfection immunity of mycobacteriophage L5: applications for genetic transformation of mycobacteria. Mol. Microbiol. **7:**407–417. http://dx.doi.org/10.1111/j.1365-2958.1993.tb01132.x.
43. **Smith MC, Thorpe HM.** 2002. Diversity in the serine recombinases. Mol. Microbiol. **44:**299–307. http://dx.doi.org/10.1046/j.1365-2958.2002.02891.x.
44. **Bibb LA, Hancox MI, Hatfull GF.** 2005. Integration and excision by the large serine recombinase phiRv1 integrase. Mol. Microbiol. **55:**1896–1910. http://dx.doi.org/10.1111/j.1365-2958.2005.04517.x.
45. **Ghosh P, Wasil LR, Hatfull GF.** 2006. Control of phage Bxb1 excision by

a novel recombination directionality factor. PLoS Biol. **4:**e186. http://dx.doi.org/10.1371/journal.pbio.0040186.
46. **Khaleel T, Younger E, McEwan AR, Varghese AS, Smith MC.** 2011. A phage protein that binds phiC31 integrase to switch its directionality. Mol. Microbiol. **80:**1450–1463. http://dx.doi.org/10.1111/j.1365-2958.2011.07696.x.
47. **Savinov A, Pan J, Ghosh P, Hatfull GF.** 2012. The Bxb1 gp47 recombination directionality factor is required not only for prophage excision, but also for phage DNA replication. Gene **495:**42–48. http://dx.doi.org/10.1016/j.gene.2011.12.003.
48. **Tanaka N, Shuman S.** 2011. RtcB is the RNA ligase component of an Escherichia coli RNA repair operon. J. Biol. Chem. **286:**7727–7731. http://dx.doi.org/10.1074/jbc.C111.219022.
49. **Botstein D, Matz MJ.** 1970. A recombination function essential to the growth of bacteriophage P22. J. Mol. Biol. **54:**417–440. http://dx.doi.org/10.1016/0022-2836(70)90119-1.
50. **Nesbit CE, Levin ME, Donnelly-Wu MK, Hatfull GF.** 1995. Transcriptional regulation of repressor synthesis in mycobacteriophage L5. Mol. Microbiol. **17:**1045–1056. http://dx.doi.org/10.1111/j.1365-2958.1995.mmi_17061045.x.
51. **Kaufmann G.** 2000. Anticodon nucleases. Trends Biochem. Sci. **25:**70–74. http://dx.doi.org/10.1016/S0968-0004(99)01525-X.
52. **Brok-Volchanskaya VS, Kadyrov FA, Sivogrivov DE, Kolosov PM, Sokolov AS, Shlyapnikov MG, Kryukov VM, Granovsky IE.** 2008. Phage T4 SegB protein is a homing endonuclease required for the preferred inheritance of T4 tRNA gene region occurring in co-infection with a related phage. Nucleic Acids Res. **36:**2094–2105. http://dx.doi.org/10.1093/nar/gkn053.
53. **Agris PF, Vendeix FA, Graham WD.** 2007. tRNA's wobble decoding of the genome: 40 years of modification. J. Mol. Biol. **366:**1–13. http://dx.doi.org/10.1016/j.jmb.2006.11.046.
54. **Sahu K, Gupta SK, Ghosh TC, 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. http://dx.doi.org/10.5483/BMBRep.2004.37.4.487.
55. **Hassan S, Mahalingam V, Kumar V.** 2009. Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv. Bioinformatics **2009:**316936. http://dx.doi.org/10.1155/2009/316936.
56. **Kunisawa T.** 2000. Functional role of mycobacteriophage transfer RNAs. J. Theor. Biol. **205:**167–170. http://dx.doi.org/10.1006/jtbi.2000.2057.
57. **Sahu K, Gupta SK, Sau S, Ghosh TC.** 2005. Comparative analysis of the base composition and codon usages in fourteen mycobacteriophage genomes. J. Biomol. Struct. Dyn. **23:**63–71. http://dx.doi.org/10.1080/07391102.2005.10507047.