Mycobacteriophages: Genes and Genomes

Graham F. Hatfull

Pittsburgh Bacteriophage Institute, Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260; email: gfh@pitt.edu

Key Words
bacteriophage, genome evolution, genomics, tuberculosis, mycobacteria

Abstract
Viruses are powerful tools for investigating and manipulating their hosts, but the enormous size and amazing genetic diversity of the bacteriophage population have emerged as something of a surprise. In light of the evident importance of mycobacteria to human health—especially *Mycobacterium tuberculosis*, which causes tuberculosis—and the difficulties that have plagued their genetic manipulation, mycobacteriophages are especially appealing subjects for discovery, genomic characterization, and manipulation. With more than 70 complete genome sequences available, the mycobacteriophages have provided a wealth of information on the diversity of phages that infect a common bacterial host, revealed the pervasively mosaic nature of phage genome architectures, and identified a huge number of genes of unknown function. Mycobacteriophages have provided key tools for tuberculosis genetics, and new methods for simple construction of mycobacteriophage recombinants will facilitate postgenomic explorations into mycobacteriophage biology.
INTRODUCTION

Mycobacteriophages are viruses that infect mycobacterial hosts. Interest in mycobacteriophages began in the late 1940s with the isolation of phages that infect *Mycobacterium smegmatis* (31, 121), followed shortly by phages that infect *Mycobacterium tuberculosis* (27). A primary motivation of these early studies was to type mycobacterial clinical isolates, which was further advanced by collecting sizable numbers of mycobacteriophages from a variety of environmental and clinical sources (37, 57, 105). The use of mycobacteriophages for typing purposes dominated the literature over the next 35 years, although important advances were made in understanding mycobacteriophage biology including the use of phage I3 as a generalized transducing phage for *M. smegmatis* (91), lysogeny in environmental and clinical strains (55, 72, 77), visualization by electron microscopy (100), and transfection of mycobacteriophage DNA (59, 114).

Mycobacteriophages emerged in the late 1980s as key players in the establishment of a facile genetic system for the mycobacteria (50). A breakthrough was established in 1987 by Jacobs et al., who used phage TM4 to construct novel shuttle phasmids that replicate as large cosmids in *Escherichia coli* and as phages in mycobacteria (53). These shuttle phasmids can be manipulated in *E. coli* using standard genetic engineering approaches and used to efficiently introduce foreign genes into mycobacteria. In the absence of other methods for direct manipulation of mycobacteriophage genomes, shuttle phasmids have proven invaluable for specialized transduction (1), transposon delivery (2, 98), and diagnostic introduction of reporter genes (51, 88). They also facilitated the use of antibiotic selectable markers through temperate phage L1 shuttle phasmids (103) and characterization of high-efficiency transformation mutants of *M. smegmatis* (104).

A notable feature of shuttle phasmid construction is that it does not require phage genomic information (52). However, realization of the full potential of mycobacteriophages for contributing to an understanding of their hosts clearly requires genomic characterization, and the first sequenced genome was that of mycobacteriophage L5 in 1993 (46). As the technologies for DNA sequencing advanced and became both quicker and cheaper, a large collection of complete mycobacteriophage genome sequences has emerged, revealing a delightfully complex, diverse, and interesting set of genomes. Seventy genome sequences are available in GenBank (Table 1) and a comparative analysis of 60 of these has been described (44).
dsDNA: double-stranded DNA

Mycobacteriophages hold considerable promise for elucidating phage diversity and evolution, gaining novel insights into the physiology and perhaps virulence of their mycobacterial hosts, and aiding the development of tools for mycobacterial genetics. In this review I focus primarily on the first of these, although the last two aspects have been greatly explored, providing insights into biofilm formation (80), cell wall composition (82, 87), tools for transposon delivery (2), reporter gene delivery (51), gene replacement (1, 118), point mutagenesis (119), single copy vectors (65), and non-antibiotic selectable markers (23), among others. Several additional reviews provide the reader with further information about mycobacteriophage genomics and applications (39–43, 75, 76). As our understanding of mycobacteriophage genomics expands, it will undoubtedly invigorate further utilities and insights.

**GENERAL PROPERTIES OF MYCOBACTERIOPHAGES**

**Mycobacteriophage Virion Morphologies**

All the characterized mycobacteriophages are double-stranded DNA (dsDNA) tailed phages belonging to the order Caudovirales. Most (61 of 70) are of the family Siphoviridae, characterized by relatively long flexible noncontractile tails, whereas nine are of the family Myoviridae, containing contractile tails (44). There is a notable absence of phages from the family Podoviridae (containing short stubby tails), although it is unclear whether their absence is due to evolutionary constraints or to physical problems in traversing the complex and relatively thick mycobacterial cell envelope.

Although the nine myoviral mycobacteriophages (Table 1) are morphologically indistinguishable, the siphoviruses show considerable variation. For example, the tail lengths vary by almost a factor of three (≈105 to ≈300 nm) and the structures at the tail tips are discernibly different in many of these phages (44). For the most part, the heads are isometric, although three—Corndog, Che9c, and Brujita—contain prolate heads, with the most extreme being Corndog, whose length-to-width ratio is almost four; the previously described but unsequenced phage R1 (106) has a prolate head similar to that of Che9c and Brujita (44). Those with isometric heads span a range of sizes, with the smallest being BPs and Halo (≈48 nm in diameter) and the largest being Bxz1 and its relatives (≈85 nm in diameter). In general, the capsid size correlates with genome size, suggesting there is a relatively constant DNA packaging density (44).

**Host Range and Host Range Determinants**

The early phage-typing studies showed that mycobacteriophages can have an almost endless variety of preferences for different bacterial hosts. Some phages (e.g., D29) have broad host ranges and infect many species of both fast-growing and slowly growing mycobacteria, including *M. smegmatis* and *M. tuberculosis* (94), whereas others (e.g., Barnyard) have very narrow preferences and infect only a single known host (94). At least one phage (DS6A) has been reported whose host range is restricted to strains composing the *M. tuberculosis* complex (10, 56), although only a partial genome sequence of this potentially extremely useful and interesting phage is available. Several phages discriminate between strains or isolates of a particular species, and we note that phage 33D differentiates between BCG strains and *Mycobacterium bovis*, and several phages have preferences for specific strains of *M. smegmatis* (C. Bowman, G. Broussard, D. Jacobs-Sera & G.F. Hatfull, unpublished observations).

For the most part, the molecular and genetic barriers to mycobacteriophage host range preferences are not known. Presumably, differentiation occurs at the cell surface due to the presence or absence of specific receptors, from the need for particular metabolic requirements after DNA has been injected into the cell, or from specific phage protection mechanisms such as immunity and restriction.
<table>
<thead>
<tr>
<th>Phage</th>
<th>Size (bp)</th>
<th>GC%</th>
<th>No. of ORFs</th>
<th>tRNA #</th>
<th>tmRNA #</th>
<th>Ends</th>
<th>Accession no.</th>
<th>Cluster</th>
<th>Origins</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bethlehem</td>
<td>52,250</td>
<td>63.3</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>AY500153</td>
<td>A1</td>
<td>Bethlehem, PA</td>
<td>45</td>
</tr>
<tr>
<td>Bxb1</td>
<td>50,550</td>
<td>63.7</td>
<td>86</td>
<td>0</td>
<td>0</td>
<td>5-frame 3i</td>
<td>AY500152</td>
<td>A1</td>
<td>Bronx, NY</td>
<td>76a</td>
</tr>
<tr>
<td>D05</td>
<td>51,621</td>
<td>63.4</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU744252</td>
<td>A1</td>
<td>Upp. St. Clair, PA</td>
<td>44</td>
</tr>
<tr>
<td>Jasper</td>
<td>50,968</td>
<td>63.7</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU744251</td>
<td>A1</td>
<td>Lexington, MA</td>
<td>44</td>
</tr>
<tr>
<td>KBG</td>
<td>53,572</td>
<td>63.6</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU744248</td>
<td>A1</td>
<td>Kentucky</td>
<td>44</td>
</tr>
<tr>
<td>Lockley</td>
<td>51,478</td>
<td>63.4</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU744249</td>
<td>A1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
</tr>
<tr>
<td>S0lon</td>
<td>49,487</td>
<td>63.8</td>
<td>86</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU826470</td>
<td>A1</td>
<td>Solon, IA</td>
<td>44</td>
</tr>
<tr>
<td>U2</td>
<td>51,277</td>
<td>63.7</td>
<td>81</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>AY500152</td>
<td>A1</td>
<td>Bethlehem, PA</td>
<td>45</td>
</tr>
<tr>
<td>Che12</td>
<td>52,047</td>
<td>62.9</td>
<td>98</td>
<td>3</td>
<td>0</td>
<td>10-base 3i</td>
<td>DQ398043</td>
<td>A2</td>
<td>Chennai, India</td>
<td>45</td>
</tr>
<tr>
<td>D29</td>
<td>49,136</td>
<td>63.5</td>
<td>77</td>
<td>5</td>
<td>0</td>
<td>9-frame 3i</td>
<td>AF022214</td>
<td>A2</td>
<td>California</td>
<td>24</td>
</tr>
<tr>
<td>L5</td>
<td>52,297</td>
<td>62.3</td>
<td>85</td>
<td>3</td>
<td>0</td>
<td>9-frame 3i</td>
<td>Z18946</td>
<td>A2</td>
<td>Japan</td>
<td>46</td>
</tr>
<tr>
<td>Pukovnik</td>
<td>52,892</td>
<td>63.3</td>
<td>88</td>
<td>1</td>
<td>0</td>
<td>10-base 3i</td>
<td>EU744250</td>
<td>A2</td>
<td>Ft. Bragg, NC</td>
<td>44</td>
</tr>
<tr>
<td>Peaches</td>
<td>51,376</td>
<td>63.9</td>
<td>86</td>
<td>0</td>
<td>0</td>
<td>10-base 3i</td>
<td>GQ303263</td>
<td>A2</td>
<td>Monroe, LA</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Bxz2</td>
<td>50,913</td>
<td>64.2</td>
<td>86</td>
<td>3</td>
<td>0</td>
<td>10-base 3i</td>
<td>AY129332</td>
<td>A2</td>
<td>Bronx, NY</td>
<td>83</td>
</tr>
<tr>
<td>Chah</td>
<td>68,450</td>
<td>66.5</td>
<td>104</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ174694</td>
<td>B1</td>
<td>Ruffsdale, PA</td>
<td>44</td>
</tr>
<tr>
<td>Colbert</td>
<td>67,774</td>
<td>66.5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>GQ303259</td>
<td>B1</td>
<td>Corvallis, OR</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Orion</td>
<td>68,427</td>
<td>66.5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398046</td>
<td>B1</td>
<td>Pittsburgh, PA</td>
<td>45</td>
</tr>
<tr>
<td>PG1</td>
<td>68,999</td>
<td>66.5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>AF547430</td>
<td>B1</td>
<td>Pittsburgh, PA</td>
<td>45</td>
</tr>
<tr>
<td>Palthonio</td>
<td>68,323</td>
<td>66.4</td>
<td>97</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>GQ303264</td>
<td>B1</td>
<td>Baltimore, MD</td>
<td>Unpublished</td>
</tr>
<tr>
<td>UncleHowie</td>
<td>68,016</td>
<td>66.2</td>
<td>98</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>GQ303266</td>
<td>B1</td>
<td>St. Louis, MO</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Qyrzula</td>
<td>67,188</td>
<td>69.0</td>
<td>81</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398048</td>
<td>B2</td>
<td>Pittsburgh, PA</td>
<td>45</td>
</tr>
<tr>
<td>Rosebush</td>
<td>67,480</td>
<td>69.0</td>
<td>90</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>AY129334</td>
<td>B2</td>
<td>Latrobe, PA</td>
<td>83</td>
</tr>
<tr>
<td>Phaedrus</td>
<td>68,090</td>
<td>67.6</td>
<td>98</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>EU816589</td>
<td>B3</td>
<td>Pittsburgh, PA</td>
<td>44</td>
</tr>
<tr>
<td>Flyer</td>
<td>69,378</td>
<td>67.5</td>
<td>103</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ3641182</td>
<td>B3</td>
<td>Pittsburgh, PA</td>
<td>Unpublished</td>
</tr>
<tr>
<td>Pipefish</td>
<td>69,059</td>
<td>67.3</td>
<td>102</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398049</td>
<td>B3</td>
<td>Pittsburgh, PA</td>
<td>45</td>
</tr>
<tr>
<td>Cooper</td>
<td>70,654</td>
<td>69.1</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398044</td>
<td>B4</td>
<td>Pittsburgh, PA</td>
<td>45</td>
</tr>
<tr>
<td>Nigel</td>
<td>69,904</td>
<td>68.3</td>
<td>94</td>
<td>1</td>
<td>0</td>
<td>Circ Perm</td>
<td>EU770221</td>
<td>B4</td>
<td>Pittsburgh, PA</td>
<td>44</td>
</tr>
<tr>
<td>Bxz1</td>
<td>156,102</td>
<td>64.8</td>
<td>225</td>
<td>35</td>
<td>1</td>
<td>Circ Perm</td>
<td>AY129337</td>
<td>C1</td>
<td>Bronx, NY</td>
<td>83</td>
</tr>
<tr>
<td>Cali</td>
<td>155,372</td>
<td>64.7</td>
<td>222</td>
<td>35</td>
<td>1</td>
<td>Circ Perm</td>
<td>EU826471</td>
<td>C1</td>
<td>Santa Clara, CA</td>
<td>44</td>
</tr>
<tr>
<td>Name</td>
<td>Species</td>
<td>Height</td>
<td>ID</td>
<td>Source</td>
<td>Location</td>
<td>ID</td>
<td>Accession</td>
<td>PhageID</td>
<td>Comment</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>--------</td>
<td>-------</td>
<td>--------</td>
<td>-------------</td>
<td>-----------</td>
<td>----------------</td>
<td>---------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Catera</td>
<td>153,766</td>
<td>64.7</td>
<td>218</td>
<td>35</td>
<td>Circ Perm</td>
<td>DQ398053</td>
<td>C1</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>ET08</td>
<td>155,445</td>
<td>64.6</td>
<td>221</td>
<td>30</td>
<td>Circ Perm</td>
<td>GQ303260.1</td>
<td>C1</td>
<td>San Diego, CA</td>
<td>Unpublished</td>
<td></td>
</tr>
<tr>
<td>LRRHood</td>
<td>154,349</td>
<td>64.7</td>
<td>227</td>
<td>30</td>
<td>Circ Perm</td>
<td>GQ303262.1</td>
<td>C1</td>
<td>Santa Cruz, CA</td>
<td>Unpublished</td>
<td></td>
</tr>
<tr>
<td>Rizal</td>
<td>153,894</td>
<td>64.7</td>
<td>220</td>
<td>35</td>
<td>Circ Perm</td>
<td>EU826467</td>
<td>C1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Scott McG</td>
<td>154,017</td>
<td>64.8</td>
<td>221</td>
<td>35</td>
<td>Circ Perm</td>
<td>EU826469</td>
<td>C1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Spud</td>
<td>154,906</td>
<td>64.8</td>
<td>222</td>
<td>35</td>
<td>Circ Perm</td>
<td>EU826468</td>
<td>C1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Myrna</td>
<td>164,602</td>
<td>65.4</td>
<td>229</td>
<td>41</td>
<td>Circ Perm</td>
<td>EU826466</td>
<td>C2</td>
<td>Upp. St. Clair, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Adiutor</td>
<td>64,511</td>
<td>59.7</td>
<td>86</td>
<td>0</td>
<td>Circ Perm</td>
<td>EU676000</td>
<td>D</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Butterscotch</td>
<td>64,562</td>
<td>59.7</td>
<td>86</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ168660</td>
<td>D</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Gumball</td>
<td>64,807</td>
<td>59.6</td>
<td>88</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ168661</td>
<td>D</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>P-loot</td>
<td>64,787</td>
<td>59.7</td>
<td>89</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398051</td>
<td>D</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Ph11</td>
<td>64,494</td>
<td>59.7</td>
<td>81</td>
<td>0</td>
<td>Circ Perm</td>
<td>DQ398047</td>
<td>D</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Troll14</td>
<td>64,618</td>
<td>59.6</td>
<td>88</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ168662</td>
<td>D</td>
<td>Silver Springs, MD</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>244</td>
<td>74,483</td>
<td>62.9</td>
<td>142</td>
<td>2</td>
<td>9-base 3í</td>
<td>DQ398041</td>
<td>E</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Cjw1</td>
<td>75,931</td>
<td>63.1</td>
<td>141</td>
<td>2</td>
<td>9-base 3í</td>
<td>AY129331</td>
<td>E</td>
<td>Pittsburgh, PA</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Kostya</td>
<td>75,811</td>
<td>62.9</td>
<td>143</td>
<td>2</td>
<td>9-base 3í</td>
<td>E816591</td>
<td>E</td>
<td>Washington, DC</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Porky</td>
<td>76,312</td>
<td>62.8</td>
<td>147</td>
<td>2</td>
<td>9-base 3í</td>
<td>E816588</td>
<td>E</td>
<td>Concord, MA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Pumpkin</td>
<td>74,491</td>
<td>63.0</td>
<td>143</td>
<td>2</td>
<td>9-base 3í</td>
<td>GQ303265.1</td>
<td>E</td>
<td>Holland, MI</td>
<td>Unpublished</td>
<td></td>
</tr>
<tr>
<td>Broome</td>
<td>58,037</td>
<td>61.1</td>
<td>105</td>
<td>0</td>
<td>10-base 3í</td>
<td>E816590</td>
<td>F1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Che8</td>
<td>59,471</td>
<td>61.3</td>
<td>112</td>
<td>0</td>
<td>10-base 3í</td>
<td>AY129330</td>
<td>F1</td>
<td>Chennai, India</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Fruitloop</td>
<td>58,471</td>
<td>61.8</td>
<td>102</td>
<td>0</td>
<td>10-base 3í</td>
<td>FJ174690</td>
<td>F1</td>
<td>Latrobe, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Lilj</td>
<td>56,852</td>
<td>61.5</td>
<td>100</td>
<td>0</td>
<td>10-base 3í</td>
<td>DQ398045</td>
<td>F1</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Pace40</td>
<td>58,554</td>
<td>61.3</td>
<td>101</td>
<td>0</td>
<td>10-base 3í</td>
<td>FJ174692</td>
<td>F1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>PMc</td>
<td>56,692</td>
<td>61.4</td>
<td>104</td>
<td>0</td>
<td>10-base 3í</td>
<td>DQ398050</td>
<td>F1</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Ramsey</td>
<td>58,578</td>
<td>61.2</td>
<td>108</td>
<td>0</td>
<td>10-base 3í</td>
<td>FJ174693</td>
<td>F1</td>
<td>White Bear, MN</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Tweety</td>
<td>58,692</td>
<td>61.7</td>
<td>109</td>
<td>0</td>
<td>10-base 3í</td>
<td>EF536069</td>
<td>F1</td>
<td>Pittsburgh, PA</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Che9d</td>
<td>56,276</td>
<td>60.9</td>
<td>111</td>
<td>0</td>
<td>10-base 3í</td>
<td>AY129336</td>
<td>F2</td>
<td>Chennai, India</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Angel</td>
<td>41,441</td>
<td>66.7</td>
<td>61</td>
<td>0</td>
<td>11-base 3í</td>
<td>EU568876.1</td>
<td>G</td>
<td>O'Hara Twp, PA</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>BPs</td>
<td>41,901</td>
<td>66.6</td>
<td>63</td>
<td>0</td>
<td>11-base 3í</td>
<td>EU568876</td>
<td>G</td>
<td>Pittsburgh, PA</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>Halo</td>
<td>42,289</td>
<td>66.7</td>
<td>64</td>
<td>0</td>
<td>11-base 3í</td>
<td>DQ398042</td>
<td>G</td>
<td>Pittsburgh, PA</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Hope</td>
<td>41,901</td>
<td>66.6</td>
<td>63</td>
<td>0</td>
<td>11-base 3í</td>
<td>GQ303261.1</td>
<td>G</td>
<td>Atlanta, GA</td>
<td>Unpublished</td>
<td></td>
</tr>
<tr>
<td>Konstantine</td>
<td>68,952</td>
<td>57.3</td>
<td>95</td>
<td>0</td>
<td>Circ Perm</td>
<td>FJ174691</td>
<td>H1</td>
<td>Pittsburgh, PA</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>Predator</td>
<td>70,110</td>
<td>56.3</td>
<td>92</td>
<td>0</td>
<td>Circ Perm</td>
<td>EU770222</td>
<td>H1</td>
<td>Donegal, PA</td>
<td>44</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Location</th>
<th>Phage Name</th>
<th>Length (bp)</th>
<th>Identity</th>
<th>Size (kb)</th>
<th>Phage Type</th>
<th>Host</th>
<th>Location</th>
<th>Identity</th>
<th>Size (kb)</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barnyard</td>
<td>Brujita</td>
<td>57,050</td>
<td>66.4</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>10-base 3í</td>
<td>1</td>
<td>1</td>
<td>Choe</td>
</tr>
<tr>
<td></td>
<td>Choe</td>
<td>57,050</td>
<td>65.4</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>10-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comdog</td>
<td>69,777</td>
<td>65.4</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>4-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Giles</td>
<td>53,746</td>
<td>67.5</td>
<td>78</td>
<td>0</td>
<td>0</td>
<td>14-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Omega</td>
<td>110,865</td>
<td>61.4</td>
<td>237</td>
<td>2</td>
<td>0</td>
<td>4-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TM4</td>
<td>52,797</td>
<td>68.1</td>
<td>89</td>
<td>0</td>
<td>0</td>
<td>10-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wildcat</td>
<td>78,296</td>
<td>56.9</td>
<td>148</td>
<td>24</td>
<td>1</td>
<td>11-base 3í</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>5,078,090</td>
<td></td>
<td>7930</td>
<td>363</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVERAGE</td>
<td></td>
<td>73,595.5</td>
<td>63.7</td>
<td>114.9</td>
<td>5.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Colored shading corresponds to genome groupings according to cluster relationships.
the normal cellular function of mpr is, or why mpr overexpression gives D29 resistance.

In many bacterial hosts, clustered regularly interspaced short palindromic repeats (CRISPRs) play roles in phage resistance (3, 116). Most sequenced mycobacterial genomes do not appear to have CRISPRs, with the exceptions being M. tuberculosis H37Rv (and related strains) and M. avium strain 104. The CRISPRs are composed of short direct repeats (21–47 bp) separated by short (∼30–50 bp) unique spacer sequences, and in the well-characterized CRISPRs the spacers have near sequence identity with phage genomic sequences, an important component for phage resistance (117). The mycobacterial CRISPR spacer sequences do not have compellingly similar counterparts in any of the sequenced mycobacteriophage genomes, consistent with the idea that many phages of these hosts remain unidentified.

**Life Cycles**

dsDNA tailed phages canonically are either temperate, forming stable lysogens at moderate frequencies (e.g., lambda), or lytic, such that all infections lead to phage growth and cell death (e.g., T4 and T7). Classification of mycobacteriophages into two such groups is, however, complex. A good example of a temperate phage is L5, which forms obviously turbid plaques from which stable lysogens immune to superinfection can be readily isolated (23); in contrast, D29 forms completely clear plaques in which virtually all host cells are killed. Genomic analysis, however, shows that D29 is a clear-plaque derivative of an L5-like temperate parent, not of a T4-like or T7-like phage (24). Of the genomically characterized phages, 12 others (the Cluster A phages; Table 1) behave similarly to L5. Most other mycobacteriophages form lightly turbid plaques, rather than clear or obviously turbid ones, and for Tweety, Giles, BPs, and Halo this reflects the ability to form lysogens at relatively low frequencies (3–5%) (78, 86, 96). Approximately one-half of the characterized mycobacteriophages (36 of 70) have an integration cassette and are candidates for forming lysogens, albeit at relatively low frequency. Phage such as Bxz1 and its relatives also form hazy plaques, although it is unclear whether the cellular survivors are uninfected cells, resistant mutants, or lysogens.

**MYCOBACTERIOPHAGE GENOMICS**

**Sequenced Mycobacteriophage Genomes**

The first completely sequenced mycobacteriophage genome was that of phage L5 (46), a temperate phage isolated in Japan (22); it is a close relative of phage L1, which shares a similar restriction pattern but does not grow at 42 °C (65). Both L5 and L1 infect fast-growing and slowly growing mycobacterial strains, although efficient infection of slow-growers by L5 requires the presence of high calcium concentrations (28). Although the sequence of L1 has not been determined, derivatives that grow at both 42 and 30 °C have been identified, followed by isolation and characterization of temperature-sensitive mutants (13, 15). The next complete genome reported was that of D29 (24), which was isolated in California from a soil sample by enrichment and infects both fast-growing and slowly growing strains, and is clearly lytic (27). D29 has considerable nucleotide sequence similarity to L5, especially in the left-most parts of the genomes that encode the virion structural genes (24). Whereas D29 forms distinctly clear plaques—perhaps more so than any other mycobacteriophage—the sequenced version is likely a recent derivative of a temperate parent, and Bowman (9) noted a mixture of plaque morphologies in his starting D29 stock; genomic comparison with L5 is consistent with this.

The third sequenced mycobacteriophage, TM4, was isolated by induction of a strain of M. avium (112). It is unclear whether the original strain was lysogenic or pseudolysogenic, since TM4 is capable of lysing it as well as M. smegmatis and M. tuberculosis (112); it does not appear to form stable lysogens in either of these
strains. Genomic analysis shows that it is distinct from L5 and D29 at the nucleotide sequence level (25), and it does not encode any known integration system or any readily identifiable phage repressor.

All the other sequenced mycobacteriophage genomes were from phages isolated over the past 20 years, and all were isolated from environmental samples using *M. smegmatis* mc²155 as a host. At the time of writing, the total number of mycobacteriophage genomes deposited in GenBank is 70 (Table 1) and a detailed comparative genomic analysis of 60 has been described (44). These phages have come from a variety of geographic locations, although about half of them were isolated from the western Pennsylvania region. The isolation of new mycobacteriophages has been greatly spurred by the development of phage discovery and genomics as an educational platform (38, 45). It would be of considerable interest to take advantage of the faster and cheaper technologies to sequence the numerous mycobacteriophages isolated in the earlier period (1950–1980)—for which detailed host range data have been reported—if these can still be recovered. We also note that the use of other mycobacterial strains for phage isolation will likely give distinct landscapes of genetic diversity to that described below for the current collection.

Overview of Genomic Diversity

The 70 sequenced mycobacteriophage genomes encompass substantial genetic diversity, and the genomic architectures are dominated by mosaic relationships. Although the overall diversity is high, it is not uniform, and any two particular phages may share either extensive nucleotide sequence similarity over the entire genome lengths with only a few base differences (e.g., phages Adjutor and PBI1), or as few as three genes whose products share greater than 25% amino acid identity (e.g., phages Barnyard and Giles) (Figure 1). Because of the mosaic nature of these genomes, many of the relationships lie between these extremes, with substantial numbers of genes shared among genomes that are not otherwise closely related.

To recognize the heterogeneous nature of genome diversity, the 70 genomes can be grouped into clusters according to their relationships to each other (Figure 1) (44). Several different methods can be used for determining the cluster assignments, including nucleotide sequence similarities and gene content analyses. For many genomes the placement into a particular cluster is simple because of extensive and clear nucleotide sequence similarity, but for other genomes it is more complex either because there is extensive but weaker similarity or because there is high nucleotide sequence similarity that extends over only a small genome segment. An arbitrary cutoff measure has been proposed that any two genomes with evident nucleotide sequence similarity spanning more than 50% of the genome lengths should be included within the same cluster (44). Using these criteria, an analysis of 60 sequenced genomes placed 55 into nine major clusters (A–I), and the remaining 5 were singleton genomes with no close relatives (44); the additional 10 genomes available in GenBank all fit within the nine major clusters (Figure 1) (Table 1). Five of these clusters can be further subdivided into subclusters, and it is anticipated that as additional genomes are sequenced new clusters will be formed (because of expected discovery of relatives of genomes that are currently singletons), and that current clusters will undergo further subdivision. The global population of mycobacteriophages would seem more likely to form a continuum of relationships, and the observed clusters may emerge from biases imposed by the isolation procedures. It is also likely that additional genomes unrelated to any of those sequenced to date remain to be discovered. Note that this clustering primarily provides a convenient framework for further analysis and does not provide an accurate portrayal of whole genome phylogenies, which involve reticulate relationships due to genomic mosaicism (64, 69, 70).

An indication that the current collection of mycobacteriophages underrepresents their full
Figure 1
Dotplot comparison of 70 sequenced mycobacteriophage genomes. Each of the 70 sequenced mycobacteriophages was concatenated into a single ∼5-Mbp sequence and compared with itself using Gepard (62). The genome order is the same as in Table 1 and the Cluster and Subcluster designations are shown above.

Diversity is provided by several prophages resident in mycobacterial genomes. Full-length prophages can be identified in the genomes of *M. avium* strain 104, *M. abscessus* (92), and *M. marinum* (108), and there are smaller prophage-like elements in *M. tuberculosis* (18, 49) and *Mycobacterium ulcerans* (107, 109). However, none of these is closely related to any of the sequenced mycobacteriophages and should be generally classified as singletons in the clustering scheme described above. The roles of any of the prophages or prophage-like sequences in virulence of their hosts are not clear, but they are of interest because many of the sequenced mycobacteriophage genomes do encode genes capable of influencing host physiology (83).

**Genome Organizations**
Mycobacteriophage genome lengths vary greatly, from 41.4 (Angel) to 164.6 kbp (Myrna), with an average length of 73.6 kbp.
Virion structure and assembly

<table>
<thead>
<tr>
<th>Terminase</th>
<th>Portal</th>
<th>Scaffold</th>
<th>Capsid</th>
<th>MTS</th>
<th>Tape measure protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>1433</td>
<td>1317</td>
<td>1406</td>
<td>1313</td>
<td>1429</td>
<td>1429</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lysis</th>
<th>Integration/ immunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td>66</td>
<td>456</td>
</tr>
</tbody>
</table>

Recombination

<table>
<thead>
<tr>
<th>Halo</th>
<th>Hope</th>
<th>BPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPME1</td>
<td>MPME1</td>
<td>MPME1</td>
</tr>
</tbody>
</table>

Figure 2
Organization of the mycobacteriophage Angel genome. The linear genome is represented by a horizontal bar with markers in kilobase pairs. Predicted genes are shown as colored boxes with the gene name shown inside the box; genes shown above the genome bar are transcribed rightward, and those below it are transcribed leftward. Each of the genes has been grouped into a phamily of related mycobacteriophage genes (44), with the Pham number designation shown above the gene. Putative gene functions are noted where known. Angel is a member of Cluster G (see Table 1), in which there are three other members, Halo, Hope, and BPs. These four genomes are similar at the nucleotide level, and differ in structure primarily by insertions of a putative mycobacteriophage mobile element (MPME). Angel contains no insertions, both Hope and BPs contain insertions of MPME1, and Halo contains an insertion of MPME2 as shown.

(Table 1). An example of genome organization is shown in Figure 2, in which the virion structure and assembly genes are arranged as an array in the left part of the genome, followed by the lysis cassette, an integration cassette, and a set of genes in the right part, some of which encode DNA replication or recombination functions, but most are of unknown function. However, there is considerable variation in genome organization and several themes emerge for different phage clusters (Figure 3). The most obvious is that all the mycobacteriophages with a siphoviral morphotype (all but Cluster C) share a syntenic group of genes encoding virion structure and assembly proteins—as seen in all siphoviruses regardless of their bacterial host and regardless of the lack of sequence similarity. For representational purposes these are shown in the left parts of the genomes (Figure 3).

Clusters F, G, and I all contain genomes with defined ends with short single-stranded DNA extensions (Table 1), and the leftmost of the structure and assembly genes (terminase) is located close to the genome end (Figure 3). In contrast, Clusters A and E, together with singletons Corndog, Giles, Omega, TM4, and Wildcat, have defined genome ends but additional genes are present between the terminase and the end (Figure 3), most of which likely do not encode virion structure and assembly functions. The number of genes varies from 4 (Cjw1; Cluster E) to 31 in the singleton Corndog, and in
Figure 3
Schematic representations of mycobacteriophage genomes architectures. The genomes of phages in the nine main clusters (A–I) and the five singleton genomes are represented by black bars with genes regions shown as colored boxes. Genes transcribed rightward are shown above the bar, and those transcribed leftward are shown below it. Putative functions of the gene blocks are represented by different colors, with the key shown at the bottom of the figure. In some clusters there is organizational variation within the cluster, and variations are given in parentheses. The genome organizations are schematic and are not drawn to scale.

Cluster A this is where the lysis genes are positioned.

Clusters B, D, and H have circularly permuted genomes, and for purposes of gene numbering and representing the genomes as linear maps, an arbitrary position close to the terminase gene is chosen as nucleotide position #1. In some genomes (e.g., Subcluster B1) this corresponds to the first base of the putative small terminase subunit gene, whereas in others it is within an upstream noncoding interval. There is a close relationship between terminase phylogeny and the nature of phage genome ends (12), and this is also observed in mycobacteriophages (44).

In many of the genomes (i.e., Clusters A, D, F, G, and I) the virion structure and assembly genes are in the canonical and largely uninterrupted order: Terminase, Portal, Protease, Scaffold, Capsid, presumed head-tail joining genes, major tail subunit, G/T tail chaperones, tape measure protein, minor tail proteins (44). Many genomes encode both small and large terminase subunits (e.g., Clusters B1, B4, E, F, G, I, Corndog, TM4), whereas in others a small terminase subunit gene has not been identified (e.g., Clusters A, B2, B3, D, H). Not all genomes encode a scaffold protein and these functions may be incorporated into the capsid subunit as they are in coliphage HK97.
The tape measure protein gene is typically the largest in the genomes of the mycobacterial siphoviruses, reflecting their rather long tails (from 107 nm in L5 to nearly 300 nm in Predator). There are, however, numerous genomes that contain additional genes in the structure and assembly gene array (Figure 3). These insertions occur at multiple locations, such as between the small and large terminase subunits (Cluster E), immediately following the major capsid subunit gene (Subcluster B1), and between the portal and protease genes (in Cluster H), and there are relatively large insertions in the singletons Corn-dog and Omega (Figure 3). The insertions in Clusters B, E, and H correspond to Holliday junction resolving enzymes (RuvC-like in Cluster B and Endo VII-like in E and H), consistent with a role for these genes in DNA packaging (36).

As noted above, in the Cluster A genomes the lysis genes are located between the terminase gene and the left end. However, this is unusual, and it is more typically located immediately downstream of the tail genes (Figures 2 and 3) and transcribed in the same direction. This is a notable departure from the lambda prototype, where the lysis functions are located close to the right end of the genome (97). Clusters A, E, F, G, and singletons Giles and Omega encode integration cassettes that are near the centers of their genomes regardless of substantial differences in genome lengths (40). In Giles the integration cassette is in an atypical location to the left of the lysis genes (78). Although genes involved in DNA replication (including DNA Pol I, Pol III, and Holliday junction resolving enzymes) and DNA metabolism (such as ThyX and ribonucleotide reductase) can be identified (Figure 3), most other genes in the siphoviral genomes are of unknown function (44).

All Cluster C mycobacteriophages have myoviral morphologies and relatively large genomes, and the virion structure and assembly genes do not appear to be organized into a well-defined array as they are in the siphoviruses. However, relatively few of the structure and assembly genes have been identified and the virion proteins are not well characterized. A striking feature of these genomes is that they encode a large number of tRNA genes (Table 1) organized into at least two large arrays. Myrna (Subcluster C2) is predicted to express 41 tRNAs, only modestly fewer than its M. smegmatis host (47 predicted tRNAs). The Subcluster C1 phages, as well as the singleton Wildcat, also encode a tmRNA gene (Table 1).

**Genome Mosaicism**

A notable feature of all bacteriophage genomes is their mosaic architectures, where each genome can be thought of as a specific assembly of individual modules (81, 83, 101). Each module may correspond to a single gene or a group of genes, and its modular nature is reflected by its location in genomes that are otherwise not closely related. The exchange of modules may have occurred relatively recently in evolutionary time, in which case the modules may retain substantial similarity at the nucleotide sequence level, or it may have occurred at more distant times, with the only remaining evidence of common descent being weak but statistically significant amino acid sequence similarity. Examples of both extremes can be found among the mycobacteriophage genomes.

An excellent example of a relatively recent exchange is seen in the Cluster B genomes (Figure 4). Cluster B genomes can be readily subdivided into four subclusters (B1–B4) such that genomes within each subcluster have high levels of nucleotide sequence similarity over their entire genome lengths, but nucleotide sequence similarity is poor between genomes of different subclusters. However, there is a ~1.9-kbp DNA sequence segment that departs from this pattern and is shared at a level of 94% nucleotide sequence similarity between phages Rosebush (a Cluster B2 member) and all six of the Cluster B1 genomes; the only other member of the B2 subcluster (Qyrzula) has a quite distinct sequence in its place (Figure 4a). Because sufficient evolutionary time has passed to allow for the accumulation of about 100 nucleotide differences between the sequences,
Recombination between Cluster B mycobacteriophages. (a) Phages Orion and PG1 are members of Subcluster B1 and are closely related at the nucleotide level (Table 1). Phages Rosebush and Qyrzula are members of Subcluster B2 and are closely related at the nucleotide level across most of their genome spans. A short portion (∼7 kbp) of the genomes is shown and aligned, with sequence similarity represented as colored shading between the pairwise genomes. The strengths of the relationships are shown according to the color spectrum, with violet representing the closest similarity. Note the segment of Rosebush that is closely related to the Subcluster B1 genomes, but not to its B2 relative Qyrzula. Genes are shown as gray boxes, with the gene name within the box, the phamily assignment above the box, and the number of phamily members in parentheses. Figure was generated using the program Phamerator (S. Cresawn, R. Hendrix & G.F. Hatfull, unpublished data). (b) Alignment of PG1, Rosebush, and Qyrzula sequences at the rightmost recombinant junction. The arrow above the sequences shows the position of the 3′ ends of genes 35 of PG1 and Rosebush; the arrows below show the 3′ and 5′ ends of Qyrzula genes 33 and 34, respectively. The box shows a region of interrupted similarity between PG1 and Qyrzula within which recombination could have given rise to the Rosebush recombinant structure.
Phamily (Pham): a group of mycobacteriophage genes related to each other as defined by BlastP and ClustalW

examination of the recombinant junctions should be interpreted cautiously. Nonetheless, at the right junction, which corresponds closely with the 3′ ends of gene 35, there is a short segment of interrupted sequence similarity between PG1 (and all of its five relatives) and Qyrzula that could have served as a site for recombination to give rise to the Rosebush structure (Figure 4b). The common sequence at the junction is not completely conserved and it is impossible to tell whether the differences have occurred subsequent to recombination, or whether they might have been present in the parent genomes (which were not necessarily Qyrzula or other known Cluster B1 phages). It has been proposed that homeologous recombination events (involving sequences that are similar but divergent) mediated by phage-encoded recombinases (such as lambda Red or the RecET systems) acting at partially conserved sequences could give rise to junctions such as these (74).

The mycobacteriophages appear to have numerous examples in which individual modules correspond to single genes, with the relationships made evident by amino acid sequence similarity (83). When the phylogenies of individual genes are determined, they are often different, revealing distinct evolutionary paths to residence in any particular phage genome. To simplify the representation of this, we have utilized phamily circles, which have the advantage of displaying all genome members used in the analysis, including those that do not contain a particular gene member of the phamily being analyzed (83). Examples are shown in Figure 5 in which both Pham 233 and Pham 471 have a member in phage Omega but have Pham members in a variety of other genomes.

In the Omega genome, genes 126 and 127 represent these two phamilies, respectively, and their distinct phylogenetic relationships strongly suggest they have evolved separately, and have been juxtaposed by a recombination event between them. This is further illustrated by examining the locations of the related pham members in other genomes. For example, Pham 233 has a member in phage Cjw1 (gene 73) that is flanked on both sides by genes unrelated to those flanking Omega gene 126. Likewise, Pham 471 has a member in phage KBG (gene 84) flanked by genes unrelated to those in Omega. This single-gene mosaicism, especially among the nonstructural genes, is a prominent feature of these genomes and underscores the dominant role of horizontal exchange processes in bacteriophage evolution.

Mechanisms for Generating Mosaic Genomes

There has been considerable speculation regarding the specific molecular mechanisms that give rise to mosaic phage genomes (47, 48). An early model suggested that short, conserved boundary sequences located at gene boundaries may serve as targets for genetic exchange (110), and such boundary sequences have been described in coliphage HK620 (17). Boundary sequences are not, however, prevalent among mycobacteriophages (or other groups of phages) and thus seem unlikely to solely account for the pervasive mosaicism. A second view is that mosaicism results from events that are primarily illegitimate or nonsequence determined. Although most of these events will be destructive,
they have the capacity to position two unrelated DNA segments together in a highly creative process. The generation of successful progeny would likely require multiple low-frequency events, coupled with selection either for gene function or for DNA segments of packagable size. The low frequency of such events would not seem to be a serious impediment in light of the dynamic nature of phage-host interactions (∼10^24 infections per second globally), the vast number of phage particles (10^31), and probable early origins extending back perhaps 3 billion years (48, 111).

A third view is that homeologous recombination plays an important role. Support for this is provided by the observation that lambda Red recombination is more proficient at recombination between divergent sequences than are host RecABCD pathways and can act at very short regions of sequence similarity (74). However, exchanges occurring at extremely short regions of sequence similarity may not be readily distinguishable from illegitimate recombination events, and exchanges at longer segments may not necessarily lead to disruptions of synteny (Figure 4). Nonetheless, the properties of phage-encoded recombination systems make them attractive for playing important roles in phage evolution, mediating exchange between short partially conserved sequences such as ribosome binding sites, transcriptional terminators, and repressor binding sites (11).

A potential caveat for a general role of lambda Red-like recombinases in generating phage mosaicism is that not all genomes obviously encode such recombination systems. In the mycobacteriophages, Clusters G, I, and Giles encode *Escherichia coli* RecET-like proteins, some of which are active in recombination (118–120); Wildcat encodes an Erf-like recombinase, and a number of other mycobacteriophages (Clusters C and E) encode RecA homologs. But recombinase genes mediating homologous exchange cannot be readily identified in the remaining 48 genomes, suggesting that they are absent or that these activities lie within the large number of genes of unknown function. It is noteworthy that high levels of recombination among TM4-derived cosmids were observed during shuttle plasmid construction (53) even though no TM4 recombination genes have been identified.

**Transposons and Other Mobile Elements**

While not all phage genomes necessarily harbor transposons or other mobile elements, they are not uncommon, and transposition is expected to contribute to genomic mosaicism. Curiously, although dozens of transposons and insertion sequences have been identified in mycobacterial genomes, none occurs in any of the sequenced mycobacteriophages (44). However, comparative genomics has revealed a novel class of mycobacteriophage mobile elements (MPMEs) that are broadly distributed among mycobacteriophage genomes (primarily in Clusters F, G, and I) but absent from other phages and mycobacterial chromosomes (96). Two main subclasses (MPME1 and MPME2) share 79% nucleotide sequence identity, although the MPME1 and MPME2 share 100% nucleotide identity within their own group (96). The MPMEs are atypically small (MPME1 is 439 bp and MPME2 is 440 bp) and generate unusual 6-bp insertions between target DNA and the left inverted repeat at the insertion site.

There is good evidence to support one additional transposon insertion. In Llij (Cluster F), gp83 is related to transposases of the IS200 family and shares 73% amino acid identity with a putative transposase from *Nocardia farcinia*. A comparison of the Cluster F genomes at the nucleotide sequence level reveals a discontinuity 96 bp upstream of the beginning of gene 83 (coordinate 48209) that likely defines the junction between the left end of a putative IS200 family element and the target. The right end is not easy to identify, and Cluster F sequence similarities are less well defined, with possible junctions either at coordinate 49731 or at coordinate 49831. The ends of other IS200 family elements form hairpin loop structures 16 bp and
6 bp from the left and right end junctions, respectively, and a plausible structure is present to the left of Llij gene 83. A second structure corresponding to the right end is less clear, raising the possibility that this element may have undergone subsequent rearrangements and may no longer be mobile.

Mycobacteriophage genomes are devoid of any clearly identifiable introns, although there are several inteins located within a variety of genes, all of which have inteinless counterparts. Five of these are terminases (encoded by phages Bethlehem, Cjw1, Kostya, Omega, and Pipefish), but the Pipefish terminase is distinct from the others in that it is circularly permuted and does not have a cos-packaging genome (44).

An intein is also present in three genes related to the Bxb1 recombination directionality factor (RDF) (gene 47) and a related intein is present in a putative nucleotidyltransferase gene in Cali (gene 3). The inteins represent highly divergent sequences, and the intein in Bethlehem gene 51 has recently been shown to represent a novel functional class (115).

---

**RDF: recombination directionality factor**

---

**MYCOBACTERIAL GENE FUNCTION AND EXPRESSION**

**Lysis**

A lysis cassette was first described for mycobacteriophage Ms6 (30, 90) and was proposed to contain five genes (Orfs 1–5). Although the complete genome sequence for Ms6 is not available, approximately 5 kbp of a 6.2-kbp sequenced segment is closely related to Cluster F phages, with Fruitloop the nearest relative (98% nucleotide identity). Of the five open reading frames (ORFs) identified, three are implicated in lysis: lysin A (Orf 2), lysin B (Orf4), and a holin (Orf 4). All the sequenced mycobacteriophages appear to encode an endolysin (lysin A), even though they are an unusual and complex group of protein sequences composed of a large number of modules assembled in multiple combinations. These modules contain many different peptidoglycan hydrolysis motifs including glycoside hydrolases, amidases, and peptidases, as well as peptidoglycan binding motifs. A direct role for these modules in lysis is demonstrated by the behavior of a lysin A–defective mutant of phage Giles, and in peptidoglycan hydrolysis by the endolysins of phages Corndog, Bxz1, and Che8 (82). The Ms6 lysin B has lipolytic activity (34) and a Giles lysin B–defective mutant forms small plaques and exhibits a lysis defect (82); the D29 lysin B protein is structurally related to cutinase-like enzymes and functions as a mycolylarabinogalactan esterase (82). Curiously, four mycobacteriophages (Che12, Rosebush, Qyzula, and Myrna) lack a lysin B homolog yet do not exhibit small plaque morphotypes like the Giles lysB mutant. An intriguing possibility is that these phages have evolved a mechanism for utilizing a host-encoded cutinase-like enzyme for this function.

**Integration and Prophage Maintenance**

Thirty-six of the sequenced mycobacteriophages (Clusters A, E, F, G, I, and singletons Giles and Omega) harbor integration cassettes composed of an integrase gene, an attP attachment site, and an RDF. With the exception of Cluster A1 phages, and phages Bxz2 and Peaches (both in Cluster A2), which encode serine integrases, most of the integrases are of the tyrosine-recombinase family. In each genome encoding a tyrosine integrase, a putative attP site can be identified owing to a short 25- to 45-bp common core region shared between the attP and attB sites in the host chromosome. Frequently, the attB core overlaps a host tRNA gene and this is observed for all characterized mycobacteriophages. In phages L5, D29, Halo, Ms6, Tweety, and Giles the attB site has been confirmed experimentally (26, 65, 78, 85, 96) and it has been predicted for Che8, Llij, PMC (which are closely related to Tweety), Che9d, Omega, Che9c, Cjw1, and 244 (86). Of the remaining phages, Fruitloop integrase (gp40) is closely related to that of Ms6 and presumably uses the same attB site; Boomer and Pace40 integrases are similar to Tweety
integrase and its relatives. A putative attP site for Brujita has yet to be identified. The M. tuberculosis prophage-like element phiRv2 is integrated into a host tRNAVal gene (49).

Integrase-mediated excisive recombination typically requires an RDF, and the best-characterized RDF for the mycobacteriophage-encoded tyrosine integrases is the 56-residue gp36 Xis of L5 (66). However, the RDF class of proteins is highly diverse (67) and only the fellow Cluster A2 phages D29, Che12, and Pukovnik encode closely related homologs. Ramsey (Subcluster F1) encodes a more distant relative (gp34), although its location adjacent to the Ramsey integrase strongly implicates it in recombination directionality control.

The functions conferring directional control in the other Cluster F phages as well as those in Clusters E and Brujita (Cluster I) remain elusive. In the Cluster G phages a putative RDF with similarity to other Xis proteins is located near the integrase gene (Figure 2), and a similar situation is observed in singletons Giles and Omega (genes 30 and 84, respectively). Che9c (Cluster I) encodes a putative RDF that is related to Giles gp30 but is located over 6 kbp away from the integrase gene.

Identification of attP and associated attB sites of serine integrases is more complicated because the common core sequence can be as short as 2–3 bp (102). However, these systems are of interest because the attB sites do not overlap tRNA genes but are within host protein-coding genes and therefore have the capacity to influence host physiology through gene inactivation or modification. A good example of this is Bxb1, which integrates into the 3’ end of the groEL1 gene of M. smegmatis (61, 80). As a result, Bxb1 lysogens are unable to form normal mature biofilms, unveiling the role of the unusual GroEL1 chaperone in the regulation of mycolic biosynthesis (80). The other Subcluster A1 phages share closely related integrases (>95% amino acid identity) and likely integrate into the same site. The Bxz2 integrase is more distantly related (27% amino acid identity with Bxb1 integrase) and integrates into a different attB site within Mmseg_5156 (86). The phage Peaches also encodes a serine integrase that is most closely related to the Bxb1 integrase (59%) but whose integration site specificity remains undetermined. The M. tuberculosis phiRv1 prophage-like element encodes a serine integrase whose attB site is unusually located within a repetitive element that provides multiple potential integration sites (7). The partial sequence of the glycopeptidolipid biosynthesis gene cluster of M. avium strain A5 shows the presence of a related serine integrase (63) that may be part of a prophage in this strain.

There are two types of RDFs associated with the mycobacteriophage-encoded serine integrases. The phiRv2 RDF is related to Xis proteins that are otherwise associated with tyrosine integrases, although its mechanism of action remains poorly defined (6). The Bxb1 RDF is not related to other known RDF proteins and was identified as the product of gene 47 through use of a genetic screen (33). Biochemical characterization shows that it is not a DNA binding protein but interacts directly with integrase-DNA complexes to promote formation of excisive synaptic complexes (32, 33).

Bxb1 gene 47 is curious in that it is conserved among mycobacteriophages encoding tyrosine integrases including L5, for which all the components required for integrative and excisive recombination are known (and do not include the L5 gp54 homolog of Bxb1 gp47) (68, 84). It presumably is involved in some function other than recombinational control, and its genomic location among DNA replication genes is consistent with a replication function. Furthermore, Bxb1 gp47 has sequence similarity to proteins of the PP2A class of phosphatases, raising the question whether phosphatase activity plays any role in recombination.

Several mycobacteriophages encode proteins containing putative nuclease domains of the ParBc superfamily, including Cluster B3, C1, and Corndog (Cluster E phages encode genes with similarity to these but which do not include ParBc domains). None of these genomes encodes an integration cassette, and it is plausible that these form lysogens in
which the genomes replicate extrachromosomally. However, none of these (or any of the 70) mycobacteriophages encodes ParA homologs and their mode of prophage maintenance, if any, remains unclear.

Gene Expression and Its Regulation

Little is known about gene expression in most mycobacteriophages. Perhaps the best understood is phage L5, where an early leftward promoter ($P_{left}$) is under the control of the phage repressor (gp71) [the closely related L1 phage has an identical repressor (99)]. The $P_{left}$ promoter is similar to $E. coli$ sigma-70 promoters, and a binding site (operator) for gp71 overlaps the promoter (11, 79). The gp71 repressor recognizes a 13-bp asymmetric sequence that is present 30 times in the L5 genome, mostly in small intergenic intervals and in one orientation relative to the direction of transcription; gp71 binding has been demonstrated for 24 of these sites (11). It is proposed that a repressor bound to these stoperator sites prevents unwanted transcripts from extending into cytotoxic phage genes during lysogeny (11). Bxb1 encodes a related repressor protein (gp69) that also binds to multiple operator and/or stoperator sites throughout the genome, although the binding site has a different consensus sequence and the phages are heteroimmune (54).

All the other Cluster A1 phages encode repressors that share at least 98% amino acid identity with Bxb1 gp69 and presumably are homoimmune. A rightward promoter, which is also repressor regulated, occurs at the right end of the Bxb1 genome (54). Multiple promoters for repressor synthesis in L1, L5, and Bxb1 are presumably required for establishment and maintenance of lysogeny (14, 54, 79), although their specific roles remain unclear. Transcriptional promoters for Ms6 lysis genes (30) are located 214 bp upstream of the first of the genes in that region, Orf 1; however, it is not clear whether this is a general feature of phages that share closely related lysis genes (in Cluster F), as the extent of sequence similarity ends approximately 60 bp upstream of Ms6 Orf 1. No late promoters for any mycobacteriophages have been identified, even though protein expression patterns suggest that these may be among the most active of all mycobacterial expression systems (25, 46). A mutant defective in late synthesis of phage L1 has been reported, but the specific genes involved are not known (21).

Other Mycobacteriophage Gene Functions

Several mycobacteriophage genes involved in DNA metabolism have been cloned and characterized. Phage L5 encodes both a thymidylate synthase (ThyX) and a ribonucleotide reductase (RNR) (gp48 and gp50, respectively), and they are expressed early in lytic growth and appear to function as a complex (5). A mutant defective in early gene expression influences expression of a proposed phage nuclease (20). Giri et al. (35) characterized an early nuclease encoded by gene 65 of phage D29 and showed that it is a structure-specific nuclease with a preference for forked structures.

Initial studies of mycobacteriophage L5 identified at least two segments of the phage L5 genome that are not well tolerated in $M. smegmatis$ and presumably encode cytotoxic proteins. Further analysis identified three cytotoxic proteins encoded by L5 genes 77, 78, and 79 (95) that prevent growth of $M. smegmatis$ when expressed and presumably interrupt specific cellular processes, although these proteins remain ill-defined. We predict that the broader mycobacteriophage collection encodes numerous additional cytotoxic proteins with considerable potential for development of antituberculosis drugs as proposed for $Staphylococcus$ phages (71).

MYCOBACTERIOPHAGE GENETIC MANIPULATION

As noted above, shuttle phasmids have been invaluable tools for constructing recombinant mycobacteriophages and for using them to deliver transposons, allelic exchange substrates, and reporter genes (1, 2, 51). However, with
M. tuberculosis is unusual among bacteria in that when linear DNA substrates are introduced by electroporation there is a high propensity for illegitimate recombination (58). Mycobacteriophages have provided two useful strategies for constructing gene replacement mutants. First, mycobacteriophage shuttle plasmids can be used to introduce allelic exchange substrate by infection, and after selection a high proportion of the progeny are the result of homologous replacement (1). Second, a mycobacterial-specific recombineering system has been developed in which RecET-like proteins encoded by mycobacteriophage Che9c are expressed to confer high levels of recombination (118). Introduction of ds-DNA or single-stranded DNA substrates into recombineering strains of M. smegmatis or M. tuberculosis provides an efficient means of generating gene replacement mutants and point mutations (118, 119). Single-stranded DNA recombineering is particularly attractive for generating isogenic strains with defined point mutations with applicability to determining the contributions of single base substitutions to the drug resistant phenotypes of multiple-drug-resistant tuberculosis and extensively drug-resistant tuberculosis clinical strains.

BRED: bacteriophage recombineering of electroporated DNA

an average mycobacteriophage genome length of over 70 kbp and packaging constraints of \( \sim 50 \) kbp in lambda particles, many mycobacteriophages are not amenable to this technology.

Bacteriophage recombineering of electroporated DNA (BRED) provides a technique for direct genetic manipulation of mycobacteriophages that takes advantage of a mycobacteria-specific recombineering system (73, 118). This recombineering approach is based on the use of the RecET-like recombination system encoded by phage Che9c, such that expression of genes 60 and 61 generates high levels of recombination in both M. smegmatis and M. tuberculosis (118, 119). In the BRED application, recombineering-proficient cells are co-electroporated with two DNA substrates; one is genomic DNA of the phage to be manipulated and the other is a short (typically 200 bp) substrate that contains the desired mutation (73). For example, a defined gene deletion can be constructed by creating a 200-bp substrate containing 100 bp homologous to each of the upstream and downstream regions (120). The mutation can be designed to minimize genetic polarity, and because recombination is efficient, there is no need to include a selectable marker or identification tag.

Following co-electroporation, plaques are recovered by plating onto lawns of a permissive bacterial host (M. smegmatis) in an infectious center configuration, i.e., by plating prior to phage replication and lysis. Each plaque therefore derives from a single cell that has taken up phage genomic DNA. Screening of 12–18 plaques by PCR typically identifies at least one plaque that is mixed, containing both wild-type and mutant alleles. Importantly, this is typically observed whether or not the gene is essential for phage growth, because if the gene is essential, then the presence of wild-type helper phage supports mutant growth in the mixed plaque. Replating for isolated plaques and screening by PCR usually identify a homogenous viable mutant (73). If a mutant is not viable, then it can be recovered with a complementing strain in which the essential gene is expressed from a plasmid (73). Because no selection is required, BRED can be used to make virtually any recombinant that is desired, including defined nonpolar deletions, insertions, point mutations, and addition of gene tags (73). BRED appears to be broadly applicable to mycobacteriophage genetic manipulation provided that plaques can be recovered by electroporation of phage genomic DNA. The BRED technology thus circumvents a major hurdle in mycobacteriophage manipulation: providing facile genetic approaches for addressing a multitude of questions in mycobacteriophage biology.

SUMMARY

In conclusion, mycobacteriophage genomics reveals that the diversity of the population is large, and that substantial parts of the population at large remain unexplored. Five of the 70 sequenced genomes have no close relatives, prophages emerging from mycobacterial genome sequencing projects are not closely related to known phages, and the diversity
of the CRISPR spacers in *M. tuberculosis* and *M. avium* genomes suggests there are many genomes yet to be discovered. With new technologies for global expression analyses and mycobacteriophage functional genomics, a new chapter in postgenomic mycobacteriophage biology is anticipated with considerable excitement.

### SUMMARY POINTS

1. Mycobacteriophage genomes are genetically highly diverse.
2. Mycobacteriophages can be grouped into clusters according to their sequence relationships.
3. Mycobacteriophage genomes are architecturally mosaic.
4. Approximately 80% of mycobacteriophage gene families are of unknown function.
5. Mycobacteriophages are sources of genetic novelty, including new classes of inteins and mobile elements.
6. BRED recombineering provides a facile and general means for constructing recombinant and mutant forms of mycobacteriophages.
7. Mycobacteriophages are rich resources for mycobacterial genetics.

### FUTURE ISSUES

1. Newly discovered mycobacteriophages isolated on a variety of different mycobacterial strains are needed to fully understand mycobacteriophage genetic diversity.
2. The potential for generating new tools for mycobacterial genetics and gaining insights into mycobacterial physiology is great and many advances await development.
3. Elucidating the function of mycobacteriophage genes will provide a fuller understanding of their biology and their evolution.

### DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

### ACKNOWLEDGMENTS

I thank my colleagues and collaborators, Roger Hendrix, Jeffrey Lawrence, Craig Peebles, Steve Cresawn, Bill Jacobs, Debbie Jacobs-Sera, and the numerous members of the Hatfull laboratory who have participated in mycobacteriophage isolation, sequencing, and analysis.

### LITERATURE CITED


   Int. J. Tuberc. Lung Dis. 3:179–84
   Microbiol. 99:223–33
   of mycobacteriophage Bxb1. Mol. Microbiol. 38:955–70
   involved in mycolic acid biosynthesis during biofilm formation in mycobacteria. Cell 123:861–73
   mycobacteriophage genomes. Cell 113:171–82
84. Pena CE, Kahlenberg JM, Hatfull GF. 1999. Protein-DNA complexes in mycobacteriophage L5 integrative 
Annual Review of Microbiology
Volume 64, 2010

Contents

Conversations with a Psychiatrist
L. Nicholas Ornston ......................................................... 1

Vaccines to Prevent Infections by Oncoviruses
John T. Schiller and Douglas R. Lowy .................................. 23

TonB-Dependent Transporters: Regulation, Structure, and Function
Nicholas Noinaj, Maude Guillier, Travis J. Barnard, and Susan K. Buchanan .......... 43

Genomes in Conflict: Maintaining Genome Integrity During Virus Infection
Matthew D. Weitzman, Caroline E. Lilley, and Mira S. Chaurushiya .................. 61

DNA Viruses: The Really Big Ones (Giruses)
James L. Van Etten, Leslie C. Lane, and David D. Dunigan ....................... 83

Signaling Mechanisms of HAMP Domains in Chemoreceptors and Sensor Kinases
John S. Parkinson .................................................................. 101

Viruses, microRNAs, and Host Interactions
Rebecca L. Skalsky and Bryan R. Cullen ...................................... 123

Basis of Virulence in Community-Associated Methicillin-Resistant Staphylococcus aureus
Michael Otto ........................................................................ 143

Biological Functions and Biogenesis of Secreted Bacterial Outer Membrane Vesicles
Adam Kulp and Meta J. Kuehn .................................................. 163

Structure, Function, and Evolution of Linear Replicons in Borrelia
George Chaconas and Kerri Kobryn ......................................... 185

Intracellular Lifestyles and Immune Evasion Strategies of Uropathogenic Escherichia coli
David A. Hunstad and Sheryl S. Justice ...................................... 203

Bacterial Shape: Two-Dimensional Questions and Possibilities
Kevin D. Young .................................................................... 223
Organelle-Like Membrane Compartmentalization of Positive-Strand RNA Virus Replication Factories
  Johan A. den Boon and Paul Ahlquist .............................................................. 241

Noise and Robustness in Prokaryotic Regulatory Networks
  Rafael Silva-Rocha and Víctor de Lorenzo ....................................................... 257

Genetic Diversity among Offspring from Archived Salmonella enterica ssp. enterica Serovar Typhimurium (Demerec Collection): In Search of Survival Strategies
  Abraham Eisenstark ............................................................................................... 277

Letting Sleeping dos Lie: Does Dormancy Play a Role in Tuberculosis?
  Michael C. Chao and Eric J. Rubin ....................................................................... 293

Mechanosensitive Channels in Microbes
  Ching Kung, Boris Martinac, and Sergei Sukharev ............................................ 313

Mycobacteriophages: Genes and Genomes
  Graham F. Hatfull .................................................................................................. 331

Persister Cells
  Kim Lewis ............................................................................................................... 357

Use of Fluorescence Microscopy to Study Intracellular Signaling in Bacteria
  David Kentner and Victor Sourjik ....................................................................... 373

Bacterial Microcompartments
  Cheryl A. Kerfeld, Sabine Heinhorst, and Gordon C. Cannon .......................... 391

Mitochondrion-Related Organelles in Eukaryotic Protists
  April M. Shiflett and Patricia J. Johnson ......................................................... 409

Stealth and Opportunism: Alternative Lifestyles of Species in the Fungal Genus Pneumocystis
  Melanie T. Cushion and James R. Stringer ....................................................... 431

How to Make a Living by Exhaling Methane
  James G. Ferry ..................................................................................................... 453

CRISPR/Cas System and Its Role in Phage-Bacteria Interactions
  Hélène Deveau, Josiane E. Garneau, and Sylvain Moineau ................................. 475

Molecular Insights into Burkholderia pseudomallei and Burkholderia mallei Pathogenesis
  Edouard E. Galyov, Paul J. Brett, and David DeShazer .................................... 495

Unique Centipede Mechanism of Mycoplasma Gliding
  Makoto Miyata ..................................................................................................... 519
Bacterial Sensor Kinases: Diversity in the Recognition of Environmental Signals
Tino Krell, Jesús Lacal, Andreas Busch, Hortencia Silva-Jiménez, María-Eugenia Guazzaroni, and Juan Luis Ramos ........................................ 539

Iron-Oxidizing Bacteria: An Environmental and Genomic Perspective
David Emerson, Emily J. Fleming, and Joyce M. McBeth .......................... 561

Fungi, Hidden in Soil or Up in the Air: Light Makes a Difference
Julio Rodríguez-Romero, Maren Hedtke, Christian Kastner, Sylvia Müller, and Reinhard Fischer .......................................................... 585

Index
Cumulative Index of Contributing Authors, Volumes 60–64 ......................... 611

Errata
An online log of corrections to Annual Review of Microbiology articles may be found at http://micro.annualreviews.org/