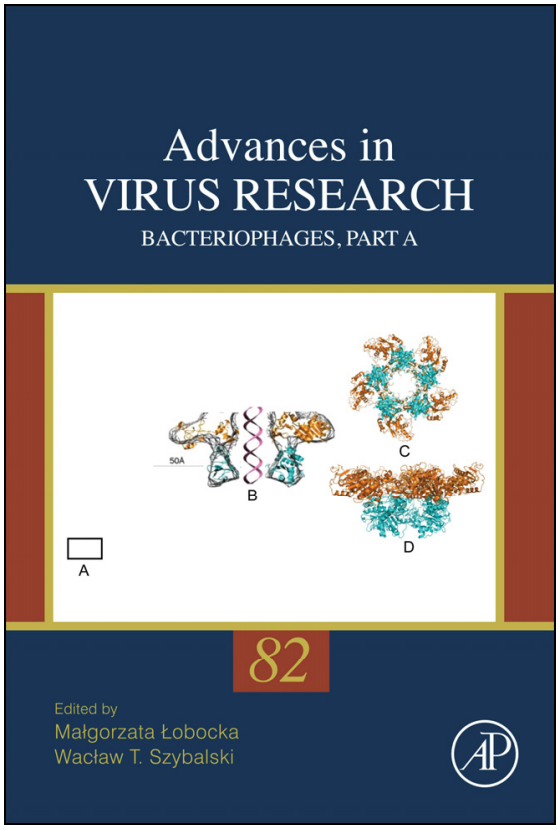
Provided for non-commercial research and educational use only.
Not for reproduction, distribution or commercial use.

This chapter was originally published in the book *Advances in Virus Research, Vol. 82*, published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who know you, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:
http://www.elsevier.com/locate/permissionusematerial

From: Graham F. Hatfull, The Secret Lives of Mycobacteriophages.
In Małgorzata Łobocka and Wacław T. Szybalski, editors:
Advances in Virus Research, Vol. 82,
Burlington: Academic Press, 2012, pp. 179-288.
ISBN: 978-0-12-394621-8
© Copyright 2012 Elsevier Inc.
Academic Press.

Author's personal copy

CHAPTER 7

The Secret Lives of Mycobacteriophages

Graham F. Hatfull

Contents

Department of Biological Sciences, Pittsburgh Bacteriophage Institute, University of Pittsburgh, Pittsburgh, Pennslyvania, USA

Advances in Virus Research, Volume 82
ISSN 0065-3527, DOI: 10.1016/B978-0-12-394621-8.00015-7

© 2012 Elsevier Inc.
All rights reserved.

Author's personal copy

Abstract

The study of mycobacteriophages provides insights into viral diversity and evolution, as well as the genetics and physiology of their pathogenic hosts. Genomic characterization of 80 mycobacteriophages reveals a high degree of genetic diversity and an especially rich reservoir of interesting genes. These include a vast number of genes of unknown function that do not match known database entries and many genes whose functions can be predicted but which are not typically found as components of phage genomes. Thus many mysteries surround these genomes, such as why the genes are there, what do they do, how are they expressed and regulated, how do they influence the physiology of the host bacterium, and what forces of evolution directed them to their genomic homes? Although the genetic diversity and novelty of these phages is full of intrigue, it is a godsend for the mycobacterial geneticist, presenting an abundantly rich toolbox that can be exploited to devise new and effective ways for understanding the genetics and physiology of human tuberculosis. As the number of sequenced genomes continues to grow, their mysteries continue to thicken, and the time has come to learn more about the secret lives of mycobacteriophages.

I. INTRODUCTION

Mycobacteriophages are viruses that infect mycobacterial hosts. Interest in these viruses first arose in the late 1940s, and more than 300 publications followed in the 1950s, 1960s, and 1970s. Many of these studies

Author's personal copy

focused on descriptions of new mycobacteriophages and their characteristics and utility in phage typing of clinical specimens. There was a significant decline in the next two decades with fewer than 100 papers published, followed by a resurgence in the early 1990s, and over 250 publications in the following two decades. This resurgence was fueled by the pioneering work of Dr. Jacobs and colleagues in using mycobacteriophages to deliver foreign DNA into mycobacteria (Jacobs, 2000; Jacobs *et al*., 1987) and by the advent of the genomics era.

The utility of exploiting mycobacteriophages to understand their pathogenic hosts—such as *Mycobacterium tuberculosis* and *Mycobacterium leprae*, the causative agents of human tuberculosis and leprosy, respectively—is enhanced by complications in growth and manipulation of their bacterial hosts (Jacobs, 1992). *M. tuberculosis* can be propagated in the laboratory with relative ease, except that it grows extremely slowly, with a doubling time of 24h, and virulent strains require biosafety level III containment. *M. leprae* cannot be grown readily under defined laboratory conditions and no simple genetic tools are available (Scollard *et al*., 2006). Mycobacteriophages multiply relatively quickly (plaques appear on a lawn of *M. tuberculosis* in 3–4days, whereas colonies take 3–4weeks to grow) and can be grown easily to high titers (Jacobs, 2000). However, isolation of new mycobacteriophages on slow-growing strains such as *M. tuberculosis* is complicated because contamination becomes a serious problem—everything else grows faster than *M. tuberculosis*. Ever since the late 1940s it has been commonplace to use relatively fast-growing saprophytic nonpathogenic strains such as *Mycobacterium smegmatis* (doubling time ~3 hours) to isolate and propagate mycobacteriophages (Mizuguchi, 1984). Some of these phages also infect *M. tuberculosis*, although many do not. However, these host preferences may be derived from host surface differences rather than metabolic restrictions on gene expression, DNA replication, packaging, or lysis (Hatfull, 2010; Hatfull *et al*., 2010).

The application of more sophisticated molecular genetic approaches has made mycobacteriophages important tools in mycobacterial genetics and has been taken advantage of in numerous ways. However, the genomic characterization of mycobacteriophages has also shown them to be enormously diverse, rendering them as fruitful subjects for addressing broader questions in viral diversity and elucidating evolutionary mechanisms (Hatfull, 2010). These dual approaches - exploration and exploitation - work well together such that key questions about mycobacteriophage biology and how they can be utilized are expanding faster than answers can be obtained. Mycobacteriophage genomics hint at a vast array of genetic and molecular secrets that await discovery, and it would seem that mycobacteriophage investigations have a very promising future—that the best is still to come.

Author's personal copy

Finally, the enormous diversity of mycobacteriophages lends them for use in an integrated research–education platform in viral discovery and genomics (Hanauer *et al.*, 2006; Hatfull *et al.*, 2006). The Science Education Alliance program of the Howard Hughes Medical Institute has facilitated implementation of mycobacteriophage discovery for freshman undergraduate students in 44 institutions in the United States since 2008, with more than 800 students engaged, hundreds of new mycobacteriophages isolated, and many dozens of genomes sequenced and analyzed (Caruso *et al.*, 2009; Pope *et al.*, 2011). This platform could be readily extended to the use of alternative bacterial hosts with the potential to have a substantial impact on the broader field of bacteriophage diversity, relieving the major limitations in the area, which are no longer in DNA sequence acquisition technologies but in obtaining individual isolates for further characterization.

This chapter discusses the current state of mycobacteriophage genomics, our current understanding of mycobacteriophage molecular biology, and the variety of ways in which mycobacteriophages have been exploited for both genetic and clinical applications. A number of other reviews on various aspects of mycobacteriophages may be useful to the reader (Hatfull, 1994, 1999, 2000, 2004, 2006, 2008, 2010; Hatfull *et al.*, 1994, 2008; Hatfull and Jacobs, 1994, 2000; McNerney, 1999; McNerney and Traore, 2005; Stella *et al.*, 2009).

II. THE MYCOBACTERIOPHAGE GENOMIC LANDSCAPE

A. Overview of 80 sequenced mycobacteriophage genomes

Consideration of mycobacteriophage diversity as revealed by their genomic characterization is a suitable starting point for this review, and a genome-based taxonomy—albeit one that is intentionally barely hierarchical—imposes a degree of order that is useful in discussing their biology. Currently, a total of 80 different phage genome sequences have been described and compared (Pope *et al.*, 2011), and as of the time of writing (January 2011) another 80 unpublished sequenced genomes are available (http://www.phagesdb.org). The discussion here is restricted primarily to the 80 published genomes listed in Table I. Mycobacteriophage genomes vary in length from 42 to 164 kbp with an average of 69.2 kbp (Table I). Genome sizes are distributed across this spectrum, although with a notable absence of phages with genomes between 110 and 150 kbp. All of the virions contain linear double-stranded DNA (dsDNA) molecules, but two different types of genome termini are observed. Approximately 60% of the phage genomes have defined ends with short single-stranded DNA (ssDNA) termini (4–14 bases), all of which have 3′ extensions. The other 40% are terminally redundant and circularly

Author's personal copy

TABLE I Genometrics of 80 sequenced mycobacteriophage genomes

Cluster	Phage	Size (bp)	GC%	#ORFs	tRNA #	tmRNA #	Ends[a]	Accession #	Origins[b]	Reference
A1	Bethlehem	52250	63.3	87	0	0	10-base 3′	AY500153	Bethlehem, PA	Hatfull *et al.*, 2006
A1	Bxb1	50550	63.7	86	0	0	9-base 3′	AF271693	Bronx, NY	Mediavilla *et al.*, 2001
A1	DD5	51621	63.4	87	0	0	10-base 3′	EU744252	Upp. St. Clair, PA	Hatfull *et al.*, 2010
A1	Jasper	50968	63.7	94	0	0	10-base 3′	EU744251	Lexington, MA	Hatfull *et al.*, 2010
A1	KBG	53572	63.6	89	0	0	10-base 3′	EU744248	Kentucky	Hatfull *et al.*, 2010
A1	Lockley	51478	63.4	90	0	0	10-base 3′	EU744249	Pittsburgh, PA	Hatfull *et al.*, 2010
A1	Skipole	53137	62.7	102	0	0	10-base 3′	GU247132	Champlin Park, MN	Pope *et al.*, 2011
A1	Solon	49487	63.8	86	0	0	10-base 3′	EU826470	Solon, IA	Hatfull *et al.*, 2010
A1	U2	51277	63.7	81	0	0	10-base 3′	AY500152	Bethlehem, PA	Hatfull *et al.*, 2006
A2	Che12	52047	62.9	98	3	0	10-base 3′	DQ398043	Chennai, India	Hatfull *et al.*, 2006
A2	D29	49136	63.5	77	5	0	9-base 3′	AF022214	California	Ford *et al.*, 1998
A2	L5	52297	62.3	85	3	0	9-base 3′	Z18946	Japan	Hatfull *et al.*, 1993
A2	Pukovnik	52892	63.3	88	1	0	10-base 3′	EU744250	Ft. Bragg, NC	Hatfull *et al.*, 2010
A2	RedRock	53332	64.5	95	1	0	10-base 3′	GU339467	Sedona, AZ	Pope *et al.*, 2011
A3	Bxz2	50913	64.2	86	3	0	10-base 3′	AY129332	Bronx, NY	Pedulla *et al.*, 2003
A4	Eagle	51436	63.4	87	0	0	10-base 3′	HM152766	Fredericksburg, VA	Pope *et al.*, 2011
A4	Peaches	51376	63.9	86	0	0	10-base 3′	GQ303263.1	Monroe, LA	Pope *et al.*, 2011
B1	Chah	68450	66.5	104	0	0	Circ Perm	FJ174694	Ruffsdale, PA	Hatfull *et al.*, 2010
B1	Colbert	67774	66.5	100	0	0	Circ Perm	GQ303259.1	Corvallis, OR	Pope *et al.*, 2011
B1	Fang	68569	66.5	102	0	0	Circ Perm	GU247133	O'Hara Twp, PA	Pope *et al.*, 2011
B1	Orion	68427	66.5	100	0	0	Circ Perm	DQ398046	Pittsburgh, PA	Hatfull *et al.*, 2006

(*continued*)

Author's personal copy

TABLE I (*continued*)

Cluster	Phage	Size (bp)	GC%	#ORFs	tRNA #	tmRNA #	Ends[a]	Accession #	Origins[b]	Reference
B1	PG1	68999	66.5	100	0	0	Circ Perm	AF547430	Pittsburgh, PA	Hatfull *et al.*, 2006
B1	Puhltonio	68323	66.4	97	0	0	Circ Perm	GQ303264.1	Baltimore, MD	Pope *et al.*, 2011
B1	Scoot17C	68432	66.5	102	0	0	Circ Perm	GU247134	Pittsburgh, PA	Pope *et al.*, 2011
B1	UncleHowie	68016	66.5	98	0	0	Circ Perm	GQ303266.1	St. Louis, MO	Pope *et al.*, 2011
B2	Qyrzula	67188	69.0	81	0	0	Circ Perm	DQ398048	Pittsburgh, PA	Hatfull *et al.*, 2006
B2	Rosebush	67480	69.0	90	0	0	Circ Perm	AY129334	Latrobe, PA	Pedulla *et al.*, 2003
B3	Phaedrus	68090	67.6	98	0	0	Circ Perm	EU816589	Pittsburgh, PA	Hatfull *et al.*, 2010
B3	Phlyer	69378	67.5	103	0	0	Circ Perm	FJ641182.1	Pitsburgh, PA	Pope *et al.*, 2011
B3	Pipefish	69059	67.3	102	0	0	Circ Perm	DQ398049	Pittsburgh, PA	Hatfull *et al.*, 2006
B4	Cooper	70654	69.1	99	0	0	Circ Perm	DQ398044	Pittsburgh, PA	Hatfull *et al.*, 2006
B4	Nigel	69904	68.3	94	1	0	Circ Perm	EU770221	Pittsburgh, PA	Hatfull *et al.*, 2010
C1	Bxz1	156102	64.8	225	35	1	Circ Perm	AY129337	Bronx, NY	Pedulla *et al.*, 2003
C1	Cali	155372	64.7	222	35	1	Circ Perm	EU826471	Santa Clara, CA	Hatfull *et al.*, 2010
C1	Catera	153766	64.7	218	35	1	Circ Perm	DQ398053	Pittsburgh, PA	Hatfull *et al.*, 2006
C1	ET08	155445	64.6	218	30	1	Circ Perm	GQ303260.1	San Diego, CA	Pope *et al.*, 2011
C1	LRRHood	154349	64.7	224	30	1	Circ Perm	GQ303262.1	Santa Cruz, CA	Pope *et al.*, 2011
C1	Rizal	153894	64.7	220	35	1	Circ Perm	EU826467	Pittsburgh, PA	Hatfull *et al.*, 2010
C1	Scott McG	154017	64.8	221	35	1	Circ Perm	EU826469	Pittsburgh, PA	Hatfull *et al.*, 2010
C1	Spud	154906	64.8	222	35	1	Circ Perm	EU826468	Pittsburgh, PA	Hatfull *et al.*, 2010
C2	Myrna	164602	65.4	229	41	0	Circ Perm	EU826466	Upp. St. Clair, PA	Hatfull *et al.*, 2010
D	Adjutor	64511	59.7	86	0	0	Circ Perm	EU676000	Pittsburgh, PA	Hatfull *et al.*, 2010
D	Butterscotch	64562	59.7	86	0	0	Circ Perm	FJ168660	Pittsburgh, PA	Hatfull *et al.*, 2010
D	Gumball	64807	59.6	88	0	0	Circ Perm	FJ168661	Pittsburgh, PA	Hatfull *et al.*, 2010
D	P-lot	64787	59.7	89	0	0	Circ Perm	DQ398051	Pittsburgh, PA	Hatfull *et al.*, 2006

Author's personal copy

D	PBI1	64494	59.7	81	0	0	Circ Perm	DQ398047	Pittsburgh, PA	Hatfull *et al.*, 2006
D	Troll4	64618	59.6	88	0	0	Circ Perm	FJ168662	Silver Springs, MD	Hatfull *et al.*, 2010
E	244	74483	62.9	142	2	0	9-base 3′	DQ398041	Pittsburgh, PA	Hatfull *et al.*, 2006
E	Cjw1	75931	63.1	141	2	0	9-base 3′	AY129331	Pittsburgh, PA	Pedulla *et al.*, 2003
E	Kostya	75811	62.9	143	2	0	9-base 3′	EU816591	Washington, DC	Hatfull *et al.*, 2010
E	Porky	76312	62.8	147	2	0	9-base 3′	EU816588	Concord, MA	Hatfull *et al.*, 2010
E	Pumpkin	74491	63.0	143	2	0	9-base 3′	GQ303265.1	Holland, MI	Pope *et al.*, 2011
F1	Ardmore	52141	61.5	87	0	0	?	GU060500	C′nty Waterford, Ireland	Henry *et al.*, 2010
F1	Boomer	58037	61.1	105	0	0	10-base 3′	EU816590	Pittsburgh, PA	Hatfull *et al.*, 2010
F1	Che8	59471	61.3	112	0	0	10-base 3′	AY129330	Chennai, India	Pedulla *et al.*, 2003
F1	Fruitloop	58471	61.8	102	0	0	10-base 3′	FJ174690	Latrobe, PA	Hatfull *et al.*, 2010
F1	Llij	56852	61.5	100	0	0	10-base 3′	DQ398045	Pittsburgh, PA	Hatfull *et al.*, 2006
F1	Pacc40	58554	61.3	101	0	0	10-base 3′	FJ174692	Pittsburgh, PA	Hatfull *et al.*, 2010
F1	PMC	56692	61.4	104	0	0	10-base 3′	DQ398050	Pittsburgh, PA	Hatfull *et al.*, 2006
F1	Ramsey	58578	61.2	108	0	0	10-base 3′	FJ174693	White Bear, MN	Hatfull *et al.*, 2010
F1	Tweety	58692	61.7	109	0	0	10-base 3′	EF536069	Pittsburgh, PA	Pham *et al.*, 2007
F2	Che9d	56276	60.9	111	0	0	10-base 3′	AY129336	Chennai, India	Pedulla *et al.*, 2003
G	Angel	41441	66.7	61	0	0	11-base 3′	EU568876.1	O′Hara Twp, PA	Sampson *et al.*, 2009
G	BPs	41901	66.6	63	0	0	11-base 3′	EU568876	Pittsburgh, PA	Sampson *et al.*, 2009
G	Halo	42289	66.7	64	0	0	11-base 3′	DQ398042	Pittsburgh, PA	Hatfull *et al.*, 2006
G	Hope	41901	66.6	63	0	0	11-base 3′	GQ303261.1	Atlanta, GA	Pope *et al.*, 2011
H1	Konstantine	68952	57.3	95	0	0	Circ Perm	FJ174691	Pittsburgh, PA	Hatfull *et al.*, 2010
H1	Predator	70110	56.3	92	0	0	Circ Perm	EU770222	Donegal, PA	Hatfull *et al.*, 2010
H2	Barnyard	70797	57.3	109	0	0	Circ Perm	AY129339	Latrobe, PA	Pedulla *et al.*, 2003
I1	Brujita	47057	66.8	74	0	0	11-base 3′	FJ168659	Virginia	Hatfull *et al.*, 2010
I1	Island3	47287	66.8	76	0	0	11-base 3′	HM152765	Pittsburgh, PA	Pope *et al.*, 2011

(*continued*)

Author's personal copy

TABLE I (*continued*)

Cluster	Phage	Size (bp)	GC%	#ORFs	tRNA #	tmRNA #	Ends[a]	Accession #	Origins[b]	Reference
I2	Che9c	57050	65.4	84	0	0	10-base 3′	AY129333	Chennai, India	Pedulla *et al.*, 2003
J	Omega	110865	61.4	237	2	0	4-base 3′	AY129338	Upp. St. Clair, PA	Pedulla *et al.*, 2003
K1	Angelica	59598	66.4	94	1	0	11-base 3′	HM152764	Clayton, MO	Pope *et al.*, 2011
K1	CrimD	59798	66.5	95	1	0	11-base 3′	HM152767	Williamsburg, VA	Pope *et al.*, 2011
K2	TM4	52797	68.1	89	0	0	10-base 3′	AF068845	Colorado	Ford *et al.*, 1998b
L	LeBron	73453	58.8	123	9	0	10-base 3′	HM152763	Allensville, NC	Pope *et al.*, 2011
Sin	Corndog	69777	65.4	122	0	0	4-base 3′	AY129335	Pittsburgh, PA	Pedulla *et al.*, 2003
Sin	Giles	53746	67.5	78	0	0	14-base 3′	EU203571	Pittsburgh, PA	Morris *et al.*, 2008
Sin	Wildcat	78296	56.9	148	24	1	11-base 3′	DQ398052	Latrobe, PA	Hatfull *et al.*, 2006
	TOTAL	5,734,561		9,013	375					
	AVERAGE	71,683	63.83	112.66	4.69					

[a] Indicates whether the genome termini are circularly permuted or if they have defined ends with the length and polarity of the ssDNA extension.
[b] The geographic location from where the phage was isolated is shown.

Author's personal copy

permuted, although the extent of the redundancy has not been determined for any of these phages (Table I). Although the average GC% content is similar to their common host *M. smegmatis*, there is substantial variation in GC% content, ranging from 56 to 69%. The implications of this are discussed further later (see Section II.D).

B. Grouping of mycobacteriophages into clusters and subclusters

Nucleotide sequence comparisons using dot plots clearly show that some mycobacteriophages are more closely related than others (Fig. 1). Although a seemingly crude approach, grouping phages according to this relatedness offers a useful and pragmatic approach that recognizes this basic level of diversity. Seventy-seven of the 80 sequenced phages can be placed in a total of 12 different clusters (A–L) with the remaining 3 considered as singletons, of which no closely related phages have yet been identified (Table I). Two of the 80 phages, Omega and LeBron, have been assigned to clusters (J and L, respectively) because they have close relatives among the sequenced but yet to be published mycobacteriophage genomes. In the case of Cluster J there are two phages in addition to Omega that form this cluster, whereas for Cluster L there are six additional phages related to LeBron. The detailed discussions that follow are constrained to just those 80 published genomes shown in Table I.

Cluster assignment is performed primarily according to recognizable nucleotide sequence similarity that spans more than 50% of the genome length with one or more other genomes (Fig. 1) (Hatfull, 2010). The advantage of using dot plot analyses for this is that it provides a method for resolving two of the most difficult scenarios that emerge: (1) when two genomes appear to have diverged substantially such that they share DNA sequence similarity over a substantial portion of their genomes, but the degree of similarity is relatively low, and (2) when two genomes share segments of DNA sequence similarity that are very similar to each other, but extend only over a relatively small portion of the genomes (i.e., <50%). In practice, relatively few such scenarios arise, and in most cases cluster assignment is straightforward. Dot plot analyses and average nucleotide identity (ANI) parameters suggest that some clusters can be further divided into subdivisions referred to as subclusters (Fig. 1) (Hatfull, 2010). Phages of different subclusters within the same cluster often share similar genome organizations and many genes are clearly orthologues as revealed by amino acid sequence comparisons of their products, but with relatively low degrees of nucleotide similarity (Fig. 1, Table I; see also Fig. 3B).

A hallmark of all or most phage genomic architectures is that they are mosaic, built from segments that have distinct evolutionary histories and

Author's personal copy

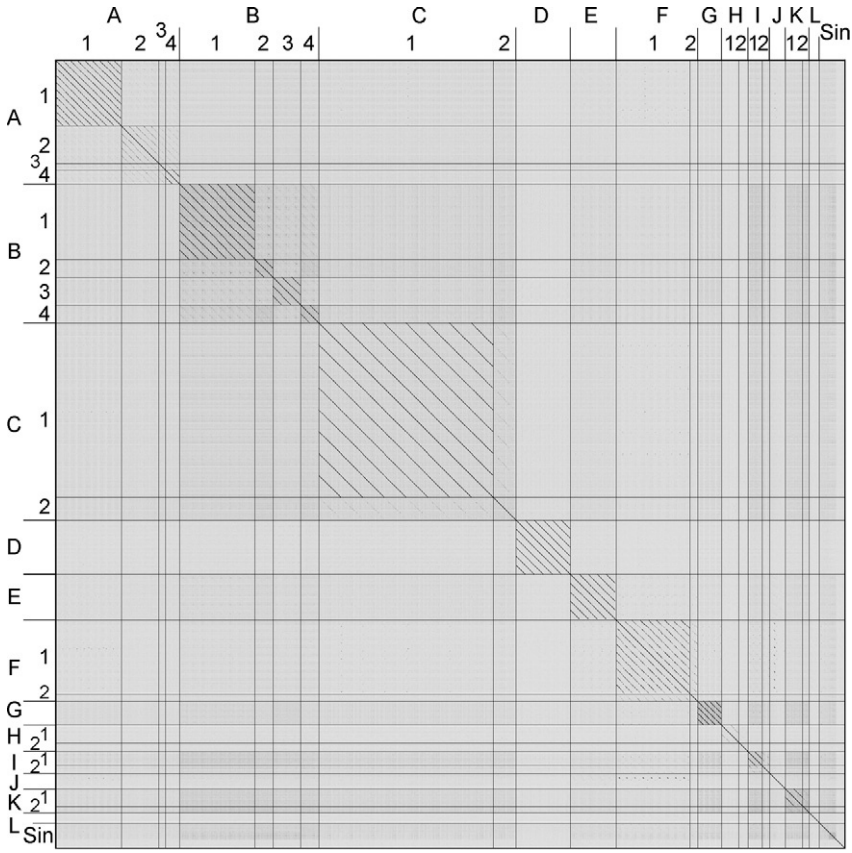


FIGURE 1 Dot plot nucleotide comparison of 80 mycobacteriophage genomes. A single FASTA-formatted file was generated containing the nucleotide sequences of all 80 sequenced and published mycobacteriophages, joined in the same order as presented in Table I. This 5.7 Mbp file was then compared to itself using the dot matrix program GEPARD (Krumsiek *et al.*, 2007). The assignment of clusters and subclusters is shown at the top and on the left.

that have been exchanged horizontally over an extended period of time; this is certainly true of mycobacteriophages (Hendrix, 2002; Hendrix *et al.*, 1999, 2000; Pedulla *et al.*, 2003). This makes any form of hierarchical classification of whole genomes difficult, and reticulate systems provide fuller and likely more accurate portrayals of their evolution (Lawrence *et al.*, 2002; Lima-Mendez *et al.*, 2008). Organization into clusters and subclusters should not be interpreted as representing any well-defined boundaries between different types of viruses, but rather a reflection of incomplete sampling of a large and diverse population of viruses occupying positions on a broad spectrum of multidimensional relationships.

Author's personal copy

As such, cluster and subcluster structures are not likely to be stable, and as more mycobacteriophages are discovered and sequenced, clusters are expected to undergo further subdivision, and differences between members of particular clusters could become reduced to that seen between subclusters. Clusters thus do not represent lineages per se, but do provide a convenient means of representing the heterogeneity of the currently sequenced phages. For example, differences between genomes within subclusters are likely to represent relatively recent evolutionary events, and a good example is the discovery of the novel MPME transposons from comparison of Cluster G phage genomes (Pope *et al.*, 2011; Sampson *et al.*, 2009).

C. Relationships between viral morphologies and cluster types

All mycobacteriophages discovered to date are tailed phages and contain dsDNA genomes (i.e., of the order Caudovirales). However, only two of the three major families of the Caudovirales are represented, and no Podoviridae—with short stubby tails—have been reported. Seventy-one of the mycobacteriophages are Siphoviridae with long flexible noncontractile tails and 9 are Myoviridae with contractile tails. All Myoviridae are in Cluster C, containing capsids approximately 80 nm in diameter and genomes 153.7 to 164.6 kbp long. Most Siphoviridae contain isometric heads—ranging from 48 to 75 nm in diameter—but several contain prolate heads, including all three members of Cluster I and the singleton Corndog (Table I)(Hatfull *et al.*, 2010; Pope *et al.*, 2011). Tail lengths of Siphoviridae vary by nearly threefold, from 110 nm to 300 (Hatfull *et al.*, 2010).

D. Relationships between GC% and cluster types

Different clusters—and in some cases subclusters—have distinctive GC% contents (Fig. 2). For example, all Cluster A genomes range between 62.3% GC% (L5) and 64.5% GC% (RedRock) (Table I), and the only other genomes that lie within this range are those in cluster E (Fig. 2). Cluster D phages are all 59.6–59.7% GC% and no other phages lie within this range. In Cluster B, the four subclusters differ somewhat in GC% with little overlap between their ranges of values, and Subclusters B1 and B4 contain genomes with the highest GC% content of any of the mycobacteriophages. At the other extreme, Cluster H and the singleton Wildcat have the lowest GC% content (56.3–57.3%). The reason why GC% should vary by cluster is not known, but an intriguing idea is that the different clusters (and perhaps in some cases subclusters) have distinct host ranges, notwithstanding that they are all capable of infecting *M. smegmatis*, a requirement of their isolation procedure; codon usage analyses are consistent

Author's personal copy

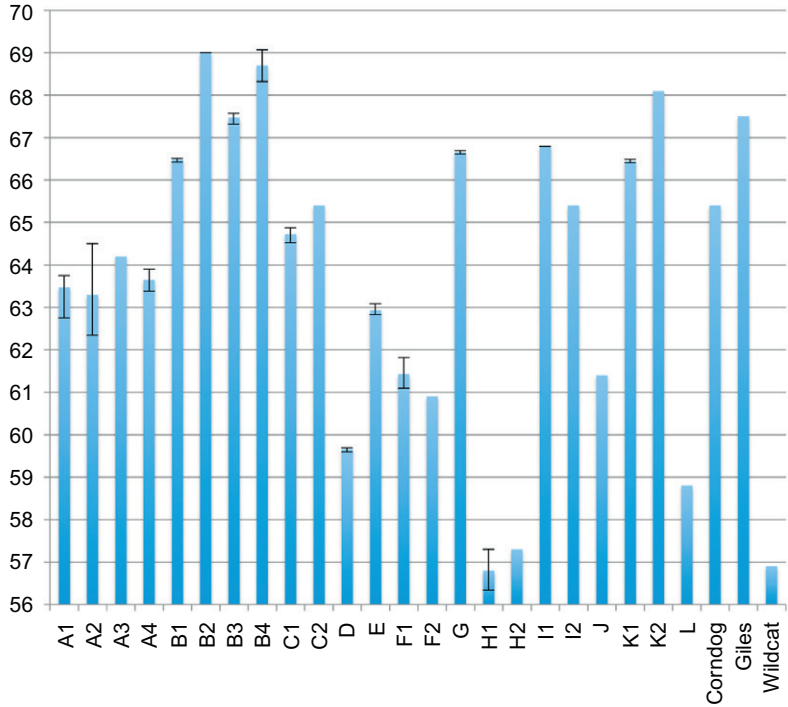


FIGURE 2 Relationships between genome GC% and cluster/subcluster types. The average GC% is shown for each cluster or subcluster of the mycobacteriophages, with variants showing extreme values within each group.

with this (Hassan *et al.*, 2009). As such, distinctions between clusters may arise from partial genetic isolation, with either host range or host availability imposing constraints on the exchange of genetic information between the genomes over a short—but evolutionary significant—time frame. Limited host range data are available for some of the phages (Rybniker *et al.*, 2006), but detailed host preferences for the larger collection of phages have yet to be determined.

E. Mycobacteriophage phamilies

Although nucleotide sequence comparisons are useful for clustering closely related phages, identification of homologues that diverged longer ago can usually only be identified by comparison of the predicted amino acid sequences. The computer program Phamerator performs automated assembly of genes into phamilies (phams) based on pairwise comparisons using both Clustal and BlastP searches using current threshold levels of 32.5% amino acid sequence identity and 10^{-50} E values, respectively

Author's personal copy

(Cresawn *et al.*, manuscript in preparation; Hatfull *et al.*, 2006; Pope *et al.*, 2011). The 80 published genomes encode a total of 9015 predicted genes, which assemble into 2343 phams of which 1106 (47.2%) are orphams (phams containing only a single gene member)(Pope *et al.*, 2011). Phamerator enables two helpful types of representation of these data. The first is the display of genome maps illustrating both regions of DNA similarity and representing individual genes according to the phamily to which they belong. The second is the use of phamily circles to display which of the component genomes contain members of any particular phamily. Phamily circles are especially useful for examining the phylogenies of adjacent genes in a genome and identifying where recombination events have adjoined genes or gene segments, each of which have distinct evolutionary histories (Hatfull *et al.*, 2006).

Database searches show that about 80% of mycobacteriophages phamilies have no identifiable homologues outside of mycobacteriophages; this high proportion of novel sequences reflects a common finding in phage genomics (Abedon, 2009; Casas and Rohwer, 2007; Comeau *et al.*, 2008; Hatfull, 2010; Hatfull *et al.*, 2006, 2010; Krisch and Comeau, 2008). As a consequence, functions of the vast majority of mycobacteriophage gene functions remain unknown. Exceptions are the approximately 10% of phamilies that are homologues of proteins with known functions and those that constitute operons of virion structural and assembly genes, that at least in phages with a siphoviral morphology have a well-conserved synteny (Hatfull *et al.*, 2010; Pope *et al.*, 2011).

F. Genome organizations

Mycobacteriophage genome organizations are well conserved among phages within clusters, and there are therefore 15 types to be considered, clusters A through L, and three singleton genomes. With the exception of Cluster C genomes, all have *siphoviral* morphologies and contain a predicted long operon of the virion structure and assembly genes, which are typically represented in the left parts of the genomes and transcribed rightward (Hatfull, 2010). In the smallest genomes (Cluster G), there are about 25 genes in this operon spanning 24 kbp, or 57% of the total genome length. In contrast, in the Cluster J phage Omega, there are 48 genes in this presumed late operon, although likely only approximately half of these have roles in virion structure and assembly genes, and the functions of most of the others are unknown (see Section III.J). However, they include a putative glycosyl transferase (gp16), a putative *O*-methyl transferase (gp17), a putative kinase (gp2), and a putative enoyl-CoA hydratase (gp4). It is not known if these are required for phage propagation or what their specific roles are.

The genomes of Clusters A, E, F, G, I, J, K, L, and all three singletons have defined ends with short (4–14 base) 3′ ssDNA extensions (Table I).

Author's personal copy

Phages in Cluster B, C, D, and H have genomes lacking defined ends, and genome sequencing data suggest that these all have circularly permuted, terminally redundant ends. In Clusters F, G, I, and K genomes, genes encoding large terminase subunits and, in some cases, genes encoding small terminase subunits can be identified close to the physical left end of the genomes, their presumed sites of action in DNA packaging. However, in other genomes, these terminase genes are displaced from the physical genome ends, with additional genes (mostly of unknown function) in the intervening space. The largest distance is in Corndog, where the large terminase subunit gene (*32*) is over 13 kbp from the left end.

Genes encoding integrases can be identified in most genomes of Clusters A, E, F, G, I, J, K, L, and singleton Giles, and are located near the center of their genomes, regardless of a span of a nearly threefold difference in genome size (Hatfull, 2006). The furthest from the midpoint is in the Cluster I phage, Brujita (39% of genome length from the left end). Putative lysis genes can be identified in all of the siphoviral mycobacteriophage genomes, although they may be located either to the left of terminase genes (as in Cluster A) or to the right of the virion structure and assembly genes (Hatfull, 2010). In Cluster C genomes, linkage of the virion structure and assembly genes is much less obvious, and the identities of relatively few have been determined. Further details of each of the genome types are discussed.

III. PHAGES OF INDIVIDUAL CLUSTERS, SUBCLUSTERS, AND SINGLETONS

A. Cluster A

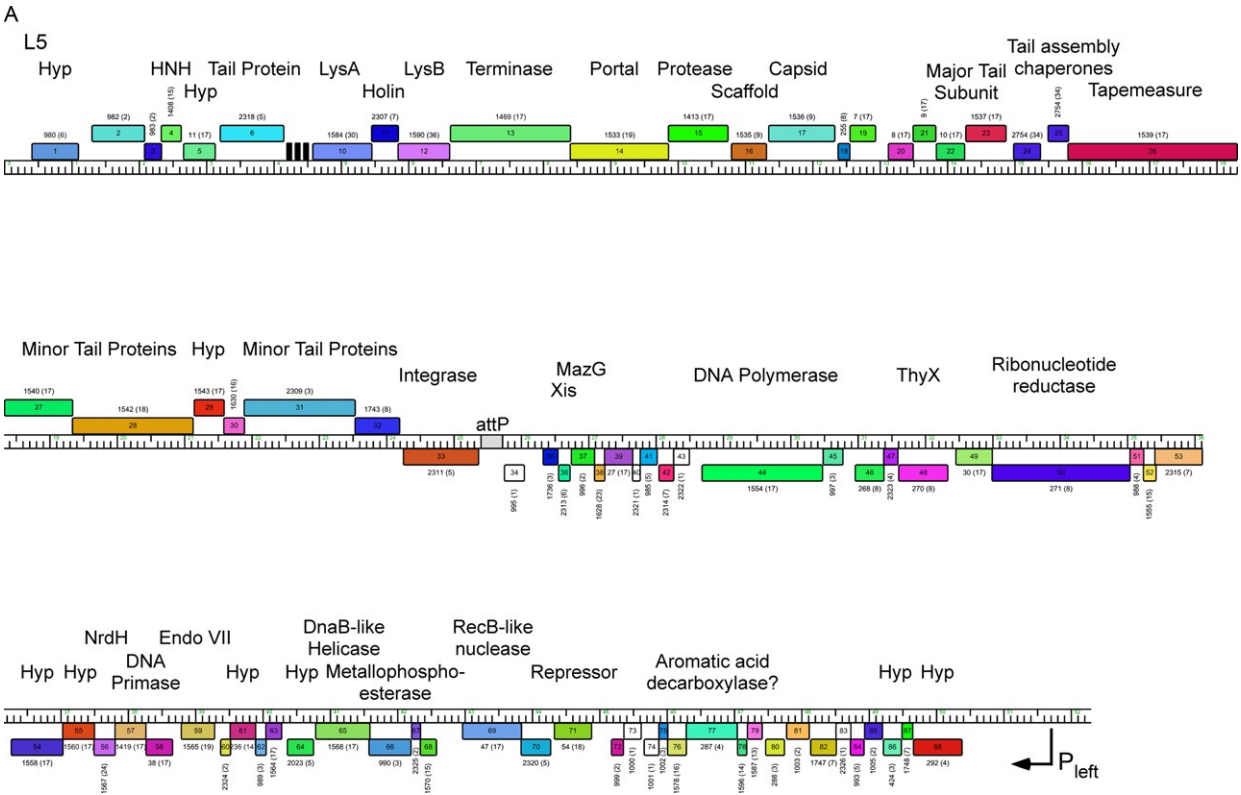
Cluster A is one of the largest clusters of mycobacteriophages, containing 17 of the 80 (21%) sequenced genomes. It is also highly diverse and currently contains four subclusters (A1–A4), although many of the unpublished genomes also belong to this cluster and are expected to expand the number of subclusters to at least six. Subclusters A1 and A2 predominate, with nine current members of A1 and five of A2 (Table I). There is no obvious geographical preference for the Cluster A phages having been isolated from two countries outside the United States and from 11 states within the United States (Table I)(Pope *et al.*, 2011). Cluster A genome lengths lie within a relatively narrow range (49,136–53,572 bp) and also occupy a narrow range of GC% (62.3–64.5%), slightly lower than that of their *M. smegmatis* host (67.4%)(Table I). Subclusters A2 and A3 phages all contain at least one tRNA gene, whereas Subclusters A1 and A4 do not. At least four of the Subcluster A2 phages infect *M. tuberculosis* efficiently [L5, D29, Che12, and Pukovnik (Fullner and Hatfull, 1997;

Author's personal copy

Gomathi *et al.*, 2007; Hatfull *et al.*, 2010; Kumar *et al.*, 2008)] but none of the phages in Subclusters A1 or A3 do so. It is not yet known if either of the Subcluster A4 phages infect *M. tuberculosis* or do so efficiently.

All Cluster A phages are either temperate or recent derivatives of temperate parents. Perhaps the best-studied member of the cluster is L5 (Subcluster A2), which was isolated in Japan in 1960 (Doke, 1960) and was the first sequenced mycobacteriophage (Hatfull and Sarkis, 1993). L5 forms evidently turbid plaques from which lysogens can be recovered readily and which are both immune to superinfection by L5 and release phage particles into culture supernatants during liquid growth. Phage L1 is a closely related temperate phage with the same restriction pattern as L5, but is naturally temperature sensitive (Doke, 1960; Lee *et al.*, 1991); its genome sequence has not yet been reported. Lysogens have also been generated for phages Bxb1, DD5, Jasper, Skipole, Solon, RedRock, Eagle, Peaches, Pukovnik, and Che12 (Kumar *et al.*, 2008; Pope *et al.*, 2011). D29, which was isolated more than 50 years ago (Froman *et al.*, 1954), forms clear plaques and kills a high proportion of infected cells. Genomic characterization suggests that it has lost a segment of approximately 3 kbp from the right end corresponding to the position of the repressor gene of L5, and D29 remains subject to L5 superinfection immunity (Ford *et al.*, 1998a). This deletion event could have occurred relatively recently, and it was noted previously that the current isolate of D29 is just one of several plaque morphotypes in the original isolate (Bowman, 1958; Hatfull, 2010). Bethlehem, KBG, Lockley, and Bxz2 also form clear plaques and fail to form lysogens (Pope *et al.*, 2011), and we speculate that the temperate nature of Cluster A phages tends to either be selected against during plaque isolation or be lost during subsequent laboratory propagation. Cluster A phages fall into three major immunity groups, which correspond closely to subclusters A1, A2, and A4 (see Section V.A.1). Lysogens of any of the temperate Subcluster A1 phages confer superinfection immunity to other A1 phages (i.e., they are homoimmune) but not to phages of other A subclusters. Similarly, phages within each of Subclusters A2 and A4 are homoimmune but heteroimmune with other Cluster A phages (Pope *et al.*, 2011). Bxz2 is currently the sole member of Subcluster A3 and does not form lysogens. However, the genome contains a putatively defective repressor gene, and because it is not subject to immunity by any of the other Cluster A phages, it likely corresponds to a derivative of a fourth distinct immunity group (Pope *et al.*, 2011). A map of the L5 genome, as a representative of Cluster A phages, is shown in Figure 3A. There are several notable features. First, the virion structure and assembly genes—several of which were identified through N-terminal sequencing of virion proteins (Hatfull and Sarkis, 1993)—are arranged in canonical order, encoding terminase, portal, protease, scaffold, capsid, major tail subunit tail assembly

Author's personal copy



Author's personal copy

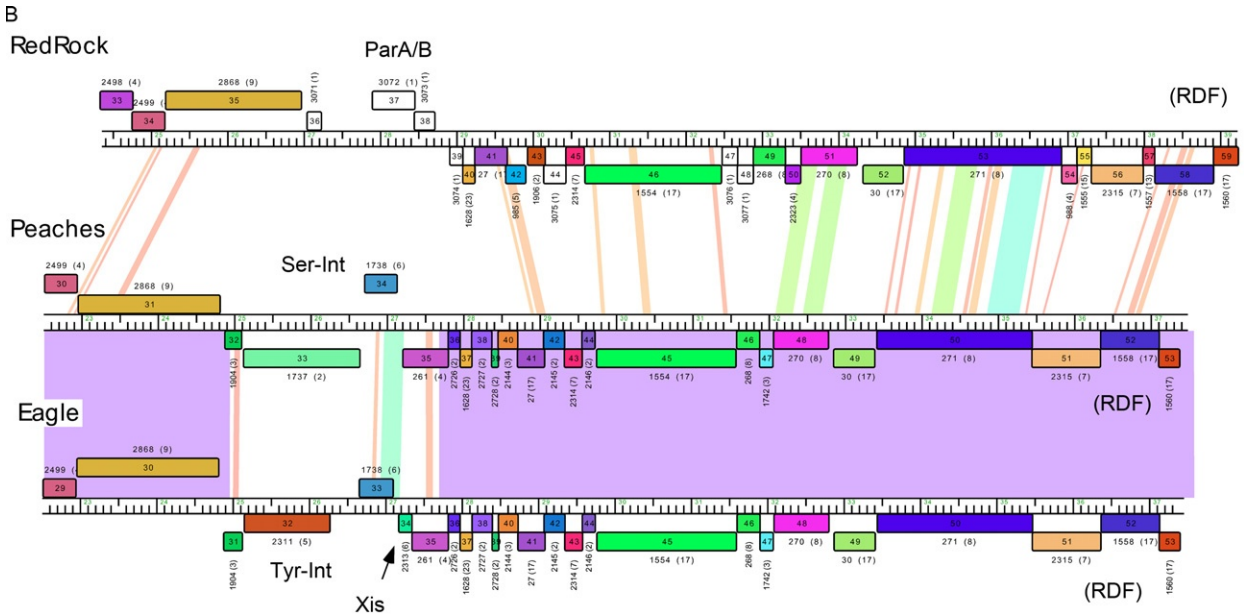


FIGURE 3 Organizational feature of Cluster A genomes. (A) Map of the L5 genome. The L5 genome (a member of Subcluster A2) is represented as a horizontal bar with markers, and the predicted ORFs are shown as colored boxes either above (rightward transcribed) or below (leftward transcribed) the genome. Gene names are shown inside the boxes, and phams to which they belong are indicated above, with the total number of pham members shown in parentheses. ORFs are color coded according to their pham memberships (i.e., all members of the same pham are the same color) and those shown in white are orphans, phams that contain only a single gene member. tRNA genes are shown as short black bars. Putative gene functions identified either experimentally or as predicted bioinformatically are shown above the genes; genes whose products match a substantial number of conserved hypothetical proteins are designated Hyp (for conserved hypothetical). Bioinformatically

Author's personal copy

chaperones, tapemeasure, and minor tail proteins (Fig. 3A), although an additional tail protein (gp6) is encoded between the terminase and the left end of the genome (Hatfull and Sarkis, 1993); no gene encoding a putative small terminase subunit has been identified. The ~6.5-kbp region between the large terminase and the left end contains three closely linked tRNA genes, as well as the lysis system. Second, the integration system is encoded in the middle of the genome, and all genes to its left are transcribed rightward and all genes to its right are transcribed leftward; the genome can therefore be split conveniently into left and right arms (Fig. 3A). Third, L5 encodes its own DNA polymerase, a Pol I-like enzyme that lacks the 5′–3′ exonuclease domain (Hatfull and Sarkis, 1993), although it is not known if it required phage DNA replication. L5 does

designated functional assignments were determined using BLASTP against the nonredundant protein database at GenBank and HHPRed (Soding, 2005). The map was generated using the program Phamerator (S. Cresawn and Graham F. Hatfull, manuscript in preparation). (B) Cassettes for genome stabilization encoding tyrosine-integrases, serine-integrases, or ParA/B partitioning functions. The figure shows representations of the central parts of three Cluster A genomes: RedRock (Subcluster A2), Peaches (Subcluster A4), and Eagle (Subcluster A4). Each genome is represented as horizontal bars with markers, and genes are represented as described (A). Maps were generated in Phamerator, and nucleotide sequence similarity between adjacent genomes is shown between them; the strength of similarity is shown according to the color spectrum, with red being the weakest and violet the strongest. Thus Peaches and Eagle are very closely related at the nucleotide level (justifying their grouping into the same subcluster, A4), except for the ~2.6-kbp central segment. RedRock is more distantly related to Peaches (and therefore to Eagle too) as shown by a sporadic, shorter, weaker segment of homology. Note that even in the absence of extensive nucleotide similarity the synteny of this region (other than in the central segment) is largely conserved, and the gene order in RedRock is similar to that in Peaches and Eagle. This can be seen by gene/pham color coding and by pham designations above the genes. Within the central region, Peaches and Eagle are different, and Peaches encodes a serine-integrase and Eagle encodes a tyrosine-integrase (gp32). The segment in Eagle that differs between the two phages also encodes the Xis protein [gp34; a member of a large highly diverse group of proteins acting as recombination directionality factors, or RDFs (Lewis and Hatfull, 2001)] that controls the directionality of integrase-mediated recombination. The central segment in Eagle contains a serine-integrase gene (*33*), but does not contain a putative RDF gene that acts with the integrase. However, the Peaches RDF is likely gp52, which is related to the known RDF of the Bxb1 serine-integrase system (Ghosh *et al*., 2006). Related proteins are encoded in all Cluster A members, even those with tyrosine-integrases such as Eagle (shown as RDF in parentheses in the figure) and likely perform additional functions in DNA replication. The Peaches serine-integrase could thus have been acquired from an Eagle-like ancestor, without concomitant RDF acquisition. In RedRock, the central segment has no integrase gene, but instead has two genes (*37* and *38*) encoding ParA and ParB functions, respectively. The RedRock prophage presumably replicates extrachromosomally and is stably maintained by these partitioning functions.

Author's personal copy

not encode its own RNA polymerase and uses the host RNA polymerase for all its transcription, being sensitive to the addition of rifampicin (Hatfull and Sarkis, 1993). Genes *48* and *50* are expressed early in lytic growth and encode flavin-dependent thymidylate synthase (ThyX) and ribonucleotide reductase functions, respectively, and the two proteins form a complex during lytic growth (Bhattacharya *et al.*, 2008). Mutants affecting two genes in L1 are implicated in the regulation of late gene expression, although locations of the mutations relative to the genome map are not known (Datta *et al.*, 2007).

There is considerable variation in the tRNA genes present in Cluster A genomes. Subclusters A1 and A4 have none, the Subcluster A3 phage Bxz2 has three ($tRNA^{asn}$, $tRNA^{trp}$, and $tRNA^{leu}$), and Subcluster A2 phages differ between one and five (Table I). Of the five tRNA genes in D29 ($tRNA^{asn}$, $tRNA^{trp}$, $tRNA^{gln}$, $tRNA^{glu}$, and $tRNA^{tyr}$), the first three are also present in L5, and the $tRNA^{glu}$ and $tRNA^{tyr}$ genes could have been lost by a simple deletion (Ford *et al.*, 1998a; Hatfull and Sarkis, 1993). Interestingly, Che12 has three similar tRNA genes, but the order of the latter two is reversed. Pukovnik and RedRock each have a single tRNA gene ($tRNA^{gln}$ and $tRNA^{trp}$, respectively), and although the RedRock $tRNA^{trp}$ is a close relative of that in L5, D29, and Bxz2, the $tRNA^{gln}$ is not closely related to the $tRNA^{gln}$ genes in L5 and D29. It has been noted that the frequencies of usage of the five amino acids corresponding to the tRNA specificities are high in D29 relative to *M. tuberculosis*, but that the specific roles of their tRNAs are uncertain because they are not well conserved among the related genomes (Kunisawa, 2000).

A leftward promoter (P_{left}) located at the right end of the genome is responsible for early expression of right arm genes and is directly under repressor control (Brown *et al.*, 1997; Nesbit *et al.*, 1995); a promoter for expression of the virion structure and assembly genes has yet to be identified. Three additional promoters are located upstream of the repressor, but it is unclear what specific roles these play (Nesbit *et al.*, 1995). Because the *attP* site is located at the 5′ side of the integrase gene, a promoter is presumably located between the *attP* crossover site and the start of the *int* gene. A detailed description of transcription patterns and a full constitution of promoter and terminator signals have yet to be elucidated for L5 or any mycobacteriophage.

A particularly intriguing feature of L5 and the other Cluster A genomes is that they contain multiple repressor binding sites—referred to as stoperators—in addition to operator sites at P_{left} (Brown *et al.*, 1997) (see Section V.A.1). In L5, there are a total of 24 sites corresponding to the 13-bp asymmetric consensus sequence 5′-GGTGGMTGTCAAG (where M is A or C) to which the repressor binds; these are located predominantly within short intergenic regions and in one orientation relative to the direction of transcription (Brown *et al.*, 1997). When one or more of

Author's personal copy

these sites is positioned between a promoter and a reporter gene, there is a repressor-dependent reduction of reporter gene activity; activity is dependent on orientation of the site relative to the direction of transcription and is amplified by multiple site insertions (Brown *et al.*, 1997). It is proposed that the repressor mediates termination of transcription rather than initiation and perhaps plays a role in ensuring transcriptional silence of phage genes in a lysogen that might otherwise be deleterious to lysogenic growth (Brown *et al.*, 1997). Several L5 genes have been implicated in cytotoxicity (Donnelly-Wu *et al.*, 1993), with strong evidence for genes *64* (Chattoraj *et al.*, 2008), *77*, *78*, and *79* (Rybniker *et al.*, 2008). All other Cluster A phages contain multiple stoperator sites with as many as 36 predicted in Jasper (Pope *et al.*, 2011). The consensus sequence is similar within each of the subclusters, but differs from subcluster to subcluster, consistent with these playing an important role—along with their cognate repressors—in determining immunity specificities.

Although there is considerable sequence diversity in Cluster A genomes, overall organizations are similar (Hatfull *et al.*, 2010). An interesting point of departure though is the use of different types of integration systems. All of the genomes in Cluster A2 encode tyrosine-integrases, whereas all A1 and A3 genomes encode serine-integrases. Interestingly, although the two phages in Cluster A4, Eagle and Peaches, are otherwise extremely similar to each other (97.5% average nucleotide identity), Eagle encodes a tyrosine-integrase and Peaches encodes a serine-integrase (Fig. 3B). The segment of DNA differing between the two genomes includes the tyrosine integrase, *attP* and the excise gene in Eagle, and the serine-integrase gene in Peaches (Pope *et al.*, 2011). We note that this genetic swap does not include the Peaches recombination directionality factor (RDF), which is presumably encoded by gene *52*, a close relative of the known RDF of Bxb1 gp47 (Ghosh *et al.*, 2006), but which is more than 9 kbp away (Fig. 3B) (see Section V.B). However, homologues of the Bxb1 RDF are present in all Cluster A genomes, regardless of whether they utilize a tyrosine-integrase or a serine-integrase, and they presumably perform additional functions, perhaps in DNA replication (Fig. 3B). Thus Peaches could conceivably have acquired the serine-integrase from an Eagle-like parent without the necessity to also acquire the RDF function (Fig. 3B). Curiously, RedRock encodes neither a tyrosine- nor a serine-integrase, but in the middle of the genome (where the integrase is located in related genomes) codes for ParA and ParB proteins, suggesting that RedRock lysogens are maintained extrachromosomally and that the ParAB system acts to provide maintenance of the prophage (Fig. 3B). This raises the question as to how an extrachromosomal RedRock prophage is replicated because its close relatives in Subcluster A2 clearly do integrate into the host chromosome and replication functions are presumably switched off. There are few clues from genome comparisons as to how

Author's personal copy

these presumed differences in replication requirements are accomplished and warrant a detailed experimental investigation. There are at least two examples of intein insertions in Cluster A phages evident from comparative genomic analyses. One of these is the intein in the Cluster A1 phage Bethlehem located within the terminase large subunit gene (*10*), which is absent from all other Cluster A terminases. Related copies of this intein are present in terminases of the Cluster L phage Omega (gp11) and Cluster E phages Cjw1 and Kostya (gp8 and gp9, respectively) and in a phamily of genes of unknown function in Cluster C phages (e.g., ET08 gp202); it is also commonly associated with most mycobacterial DnaB-like helicases. Interestingly, the Bethlehem gp10 intein utilizes a novel mechanism of protein splicing (Tori *et al.*, 2009). The second type of intein is present in Cluster A1 phages Bethlehem gp51, KBG gp53, and U2 gp50, which are homologues of the RDF protein that controls the directionality of Bxb1 serine-integrase-mediated site-specific recombination (see Section V.B.2)(Ghosh *et al.*, 2006). Relatives of this intein are also present in Cluster C phages Cali gp3, ET08 gp3, and LRRHood gp3 that encode putative nucleotidyltranferases (see Section III.C).

Cluster A virions are all siphoviral in their morphology and contain isometric heads attached to long flexible tails. The lengths of their tails are relatively short (~115 nm) compared to other mycobacteriophages (Hatfull *et al.*, 2010), and a close correlation exists between tail length and length of the gene encoding the tapemeasure protein (Pedulla *et al.*, 2003). Between the major tail subunit gene (*23*) and the tapemeasure protein gene (26) are two open reading frames (*24* and *25*) that are functionally analogous to the *G* and *T* genes of phage Lambda (Levin *et al.*, 1993) and are expressed via a −1 programmed translational frameshift (Xu *et al.*, 2004); these are thought to act as tail assembly chaperones. Interestingly, although the capsid subunits are not closely related to the well-studied phage HK97 capsid subunit, L5, D29, and Bxb1 (and presumably all other Cluster A phages) share the property of wholesale covalent cross-linking of capsid subunits (Ford *et al.*, 1998a; Hatfull and Sarkis, 1993; Mediavilla *et al.*, 2000; Popa *et al.*, 1991).

Genome comparison of Cluster A phages shows that Bethlehem contains a segment including genes *71* and *72* that is absent from other Subcluster A1 phage genomes. The gene products are distantly related to genes found in Omega (*21*, *22*) postulated to be part of an IS110-like transposon (see Section III.J). Bethlehem therefore likely carries a distantly related member of this poorly characterized transposon family.

B. Cluster B

Cluster B contains 15 phages whose genomes range from 67,118 to 70,654 bp. There are four subclusters, B1–B4, and there are a number of notable differences among them. Cluster B virions contain a linear genome with

Author's personal copy

terminally redundant and circularly permuted ends (Table I). The left ends of B1 genomes are arbitrarily designated as the first base of the putative small terminase gene, but in Subclusters B2, B3, and B4, additional genes are closely linked with intergenic spaces to the left of putative terminase large subunit genes, and thus the left end of the genome is arbitrarily designated at the first noncoding intergenic gap encountered to the left of the terminase gene. Cluster B genomes could use specific packaging sites near the terminase gene to initiate headful packaging, but none have been identified. Cluster B phages typically form plaques that are neither clear nor evidently turbid, but have a somewhat hazy appearance. However, stable lysogens have not been reported for any Cluster B phage, and they behave as lytic rather than temperate phages. The genomes provide few clues as to their life styles, and none encode identifiable integrases, transposases, or partitioning functions. Also, none encode recognizable repressors, although these are generally diverse at the sequence level, and it is noteworthy that the L5 repressor (gp71) has no close relatives outside of the mycobacteriophages; repressors can thus be overlooked easily. Like Cluster A phages, Cluster B phages encode their own DNA polymerase—also a Pol I-like enzyme—as well as a putative primase/helicase protein. The genome organization of the Cluster B representative, Rosebush, is shown in Figure 4. The virion structure and assembly genes are shown in the left part of the genome and transcribed rightward; the long tapemeasure protein gene is a striking feature because of its length (5.6 kbp), reflecting the long tail (235 nm) in Rosebush and other Cluster B virions (Hatfull *et al.*, 2010). However, there are several notable features of this presumed operon. First, it is interrupted by a number of genes of unknown function, transcribed in both forward and reverse directions (Fig. 4). The leftward-transcribed genes *8* and *23* both have homologues outside of Cluster B, raising confidence in their annotation and identification. There are also four rightward-transcribed genes (*9–12*) between the terminase (*7*) and portal (*13*) genes; gp8 is related to Holliday Junction resolving RuvC-like proteins, but the others are of unknown function. Second, there are five predicted rightward-transcribed genes between the major tail subunit gene (*21*) and the tapemeasure protein gene (*29*; and two transcribed leftward) instead of the more typical pair of genes as seen in L5 (Fig. 3A). Moreover, the two genes in L5 (*24* and *25*) encoding tail assembly chaperones are expressed via a programmed translational frameshift, a highly conserved feature in virtually all dsDNA-tailed viruses, especially those with siphoviral morphotypes (Xu *et al.*, 2004). Cluster B phages appear to be notable exceptions, and although one or more of genes *24–28* could perhaps act as tail assembly chaperones, there is no evidence supporting expression via frameshifting. We note that none of these have mycobacteriophage-related proteins outside of Cluster B phages. Two of the Cluster B3 phages, Pipefish and Phlyer, contain intein insertions in their terminase large subunit genes (gp6).

Author's personal copy

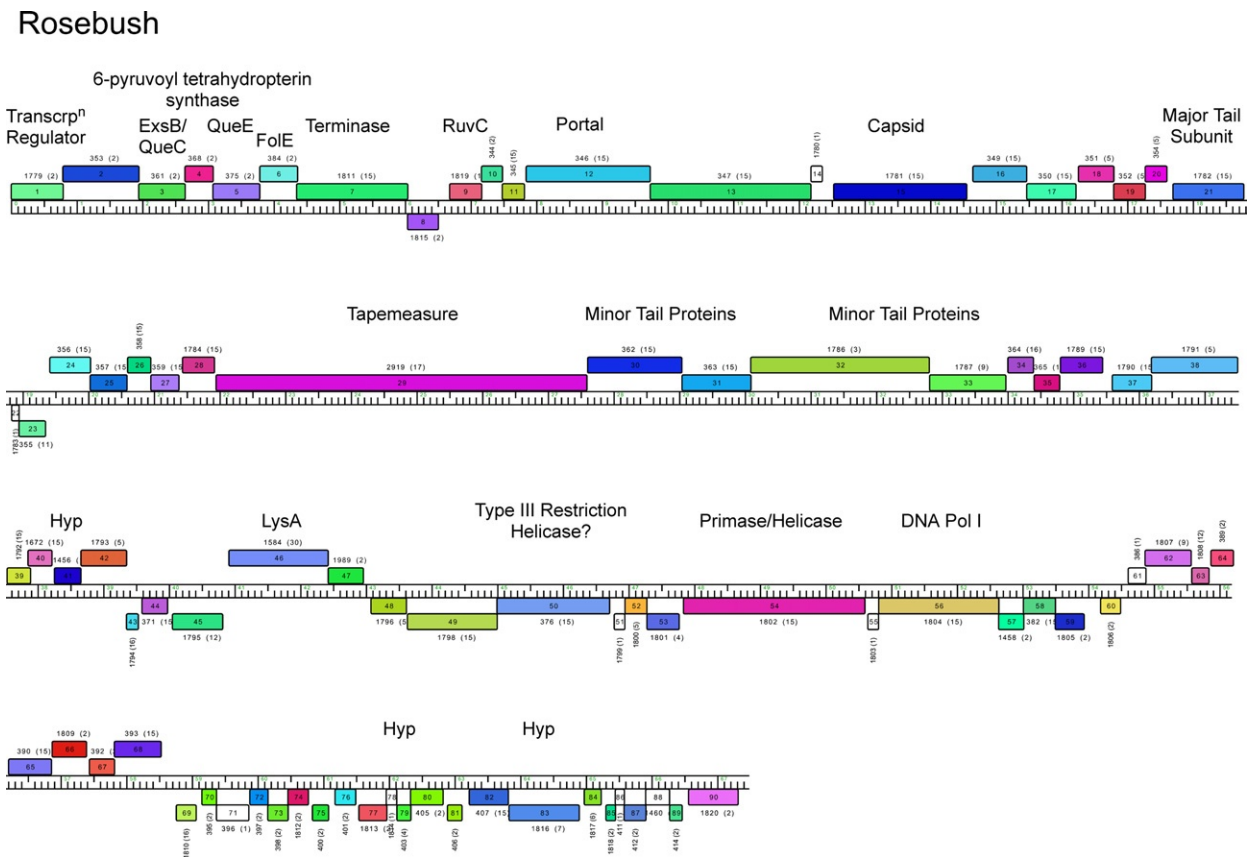


FIGURE 4 Map of the phage Rosebush genome, a member of Subcluster B2. See Figure 3A for further details on genome map presentation.

Author's personal copy

The extent of the Rosebush virion structure and assembly operon is unclear, although closely linked open reading frames continue after the tapemeasure protein gene through to gene *42* (there is a small noncoding gap between genes *36* and *37*), all or many of which are likely to encode minor tail proteins (Fig. 4). Genes to the right of gene *42* are arranged in five putative operons containing genes *43–45*, *46–47*, *48–60*, *61–68*, and *69–90*. The vast majority of these 46 genes are of unknown function, although the *48–60* operon includes genes encoding DNA replication functions. Rosebush gene *46* corresponds to lysin A, and it is plausible that *47* encodes a holin required for the delivery of Lysin A to its peptidoglycan substrate. Rosebush and its Subcluster B2 associate Qyrzula are notable in that they are among the few mycobacteriophages that do not encode a Lysin B protein (Payne *et al.*, 2009); Subcluster B1, B3, and B4 phages all encode a Lysin B protein. Subcluster B2 phages (Rosebush and Qyrzula) have an intriguing set of six genes (*1–6*) located between the terminase gene (*7*) and the arbitrarily designated genome left end (Figs. 4 and 5). Genes *3–6* are predicted to encode QueC, QueD, QueE, and QueF proteins, respectively, strongly implicating them in the biosynthesis of queuosine from GTP; HHPred analysis (Soding, 2005) predicts that gp2 is a queuine-tRNA-ribosyltransferase (Fig. 5). Queuosine is a modified base found commonly in bacterial tRNAs in the first anticodon position, although known queuosine biosynthetic genes are absent from mycobacterial genomes and the presence of queuosine as a tRNA modification has not been reported. Neither Rosebush nor any other Cluster B phage encodes its own tRNAs, and we therefore predict that in Rosebush (and Qyrzula) infections, that host tRNAs are modified with Queuosine; whether this alters the specificity of the translational apparatus or simply enhances the efficiency of translation is not known. Rosebush gene *1* is predicted to encode a transcription regulator but HHPred predicts relationships to ParB, KorB, and SopB proteins involved in chromosome partitioning; the proximity to the Queuosine biosynthetic operon suggests that gp1 could be involved in regulating this process.

The architectures of Subcluster B1, B3, and B4 genomes are similar to those of Rosebush and Qyrzula (Subcluster B2), including the long putative operon containing the virion structure and assembly genes and the five predicted operons to its right. There is an interesting difference within the structural gene organization in that—like Subcluster B2—Subclusters B1, B2, and B4 also encode homologues of RuvC (Rosebush *9*), but is transcribed in the opposite direction (Fig. 5). Maintenance of the RuvC gene in this location—notwithstanding the genetic rearrangements—is consistent with this playing an important role in virion assembly, perhaps in resolving any residual Holliday Junctions in replicated DNA molecules that otherwise would not be packaged. We note that HJ resolvases are also present within the virion structure gene operons of some other mycobacteriophages (e.g., Cluster E phages 244, Pumpkin and Porky).

Author's personal copy

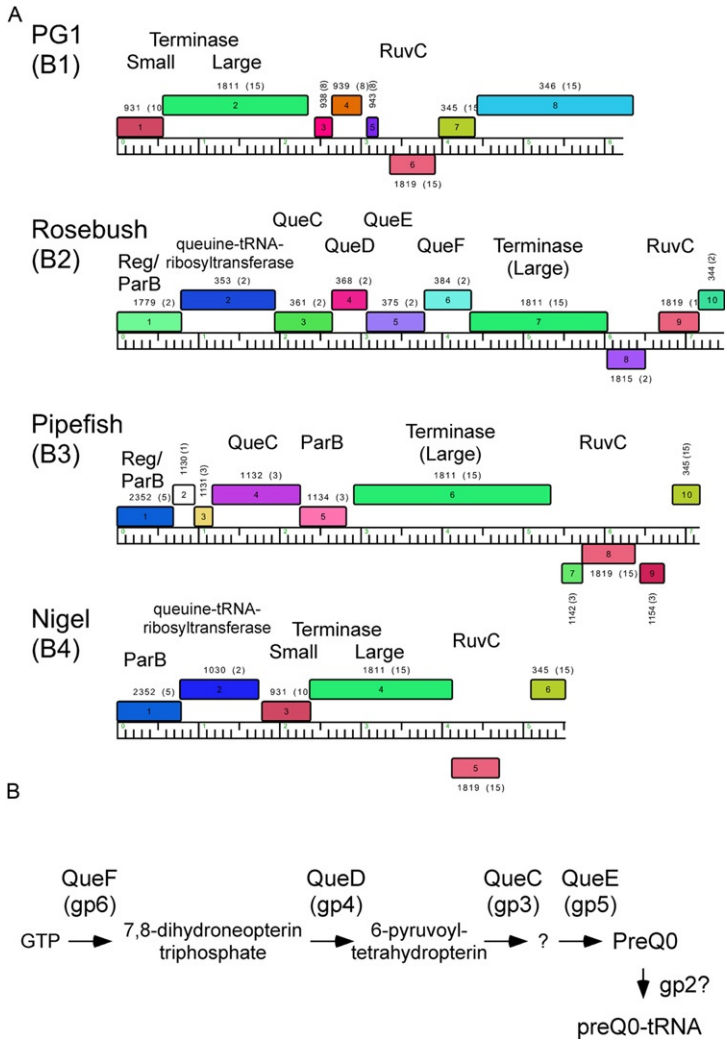


FIGURE 5 An unusual group of genes in Cluster B genomes. (A) Genome maps of segments of phages PBI1, Rosebush, Pipefish, and Nigel, representing Subclusters B1, B2, B3, and B4, respectively, are shown. Segments correspond to the extreme left ends of the genomes as they are standardly represented, although these genomes are circularly permuted and terminally redundant, the left end is arbitrarily designated and is not a physically defined left end. However, genes to the left (i.e., represented at the extreme right end of the genomes) are not evidently related to any of the functions described here. Five of the six genes to the left of the terminase large subunit gene of Rosebush (*7*) are predicted to be involved in queuosine biosynthesis, which presumably is used for tRNA modification (see text), as well as a predicted regulator with ParB-like features (gene *1*). PG1 contains none of these but has a putative terminase small subunit gene (*1*). Pipefish and Nigel have a subset of the Rosebush functions, including some but not all of

Author's personal copy

Notable differences exist in the region to the left of terminase genes, where putative queuosine biosynthetic genes are located in Subcluster B2 phages (Fig. 5). Subcluster B1 and B4 phages encode a terminase small subunit gene in this location, but Subcluster B2 and B3 phages appear to lack this function. Subcluster B4 phages (e.g., Nigel), however, encode two additional rightward-transcribed proteins (gp1 and gp2), and gp1 has similarities to ParB proteins, mirroring the putative function of Rosebush gp1 (although the two proteins are not obviously homologues). Interestingly, gp2 is predicted by HHPred to encode a queuine-tRNA-ribosyltransferase, although the protein is not obviously related to Rosebush gp2, and the Subcluster B4 phages do not encode other queuosine biosynthetic genes (Fig 5). The subcluster B3 phage Pipefish has five genes upstream of the terminase large subunit gene (*5*) (Phaedrus has only four of these), and it is noteworthy that Pipefish gp1 is a distant relative of Rosebush gp1, and Pipefish gp4 is likely a QueC-like protein, but without significant sequence similarity to Rosebush gp3 (Fig. 5). Pipefish gp5 is related to ParB-like proteins, although not related with significant sequence similarity to Rosebush gp1. The roles of the gene encoded to the left of the terminase genes in the Cluster B phages is therefore a substantial mystery, especially as there is great sequence divergence but with conservation of some common functions.

C. Cluster C

Cluster C currently contains nine phages divided into two subclusters, C1 and C2; C2 contains just a single phage, Myrna. All form plaques that are not completely clear, but also do not form stable lysogens, and their genomes do not encode any recognizable features of temperate phages. In general, the eight Cluster C1 phages are very similar to each other, while Myrna differs in a variety of ways. The C1 genomes are all relatively long but similarly sized (153.7–156.1 kbp) and Myrna is substantially larger (165.6 kbp), the largest of all the sequenced mycobacteriophage genomes. All of the Cluster C phages have myoviral morphologies with 80-nm-diameter isometric heads and modest length (85 nm) contractile tails. Cluster C genomes are circularly permuted and terminally

the genes. (B) Biosynthetic pathways for queuosine biosynthesis. QueF (Rosebush gp6) is a member of the family of GTP cyclohydrolases (which also includes FolE) and converts GTP to 7,8-dihydroneopterin triphosphate. This is converted to 6-pyruvoyltetrahydropterin by QueD (Rosebush gp4), which is then converted to PreQ0 by QueC and QueE (Rosebush gp3 and gp5, respectively). PreQ0 is transferred to a tRNA substrate by a queuine-tRNA-ribosyltransferase (Rosebush gp2). None of these phages appears to encode enzymes that would further process the PreQ0-tRNA to Q-tRNA.

Author's personal copy

redundant, although some (such as Bxz1) are unusual in that there is a long run of G-C residues in the genome through which sequencing reactions typically fail (Pedulla *et al*., 2003). The left end of Bxz1 is arbitrarily designated as being the first unique base position after the G-string, and the other Cluster C phage genomes are represented and numbered accordingly.

The genome organization of the Cluster C representative Bxz1 is shown in Figure 6. It differs from all other mycobacteriophages in that few of the virion structure and assembly genes have been defined and it is not obvious that they are organized syntenically with respect to the genomes of Siphoviridae. However, Bxz1 gp124 and gp125 are similar to major structural proteins of mycobacteriophage I3 (Ramesh and Gopinathan, 1994), and gp129 is a putative tapemeasure protein, located immediately downstream of two genes (*127* and *128*) predicted to be expressed via a programmed translational frameshift (Fig. 6). Bxz1 gp135 and gp143 have similarity to other mycobacteriophage minor tail proteins, gp133 has features suggesting it is a plausible minor tail protein, and gp137 is related to Baseplate J proteins. Thus the region encoding genes *124–143* likely corresponds to a set of genes involved in tail structure and assembly; we note that Bxz1 gp114 has similarities to some other mycobacteriophage tail proteins (Fig. 6). The location of the putative major capsid subunit is unclear, although gp112 has similarity to phage head decoration proteins.

Bxz1 and its fellow Cluster C phages have a variety of genes whose putative functions suggest enticing aspects of its biology. For example, Bxz1 gp29 is a Band-7 family protein with two strongly predicted membrane-spanning segments at its N terminus (Fig. 6). Is this protein associated with host membranes during phage replication and, if so, what is its function? Is it possible that there are membranes associated with the virion itself and that gp29 is a virion protein? Answers to these questions remain unresolved. Bxz1 also encodes two putative glycosyltransferases and a galactosyltransferase protein, and gp230 is related to Ro proteins (Fig. 6). These phages also encode a large number of tRNA genes (>30; Table I)(Sahu *et al*., 2004) as well as a tmRNA gene, a putative initiation factor (gp200), and a peptidyl-tRNA hydrolase (gp164), suggesting substantial modifications to the host translational machinery (Fig. 6). They encode a large DNA Pol III α subunit and DnaB/C proteins implicated in replication initiation and synthesis. In addition to RecA (gp220), Bxz1 also encodes a RusA-like Holliday junction resolvase (gp205). Many Cluster C genomes contain one or more intein insertions. Phage ET08 has a total of five inteins, more than any other mycobacteriophage genome. One of these is located in the gp3 putative nucleotidyltranferase gene, and Cali gp3 and LRRHood share this intein, as do the RDF proteins of some Cluster A1 phages (see Section III.A). A second is in ET08 gene *79* and

Author's personal copy

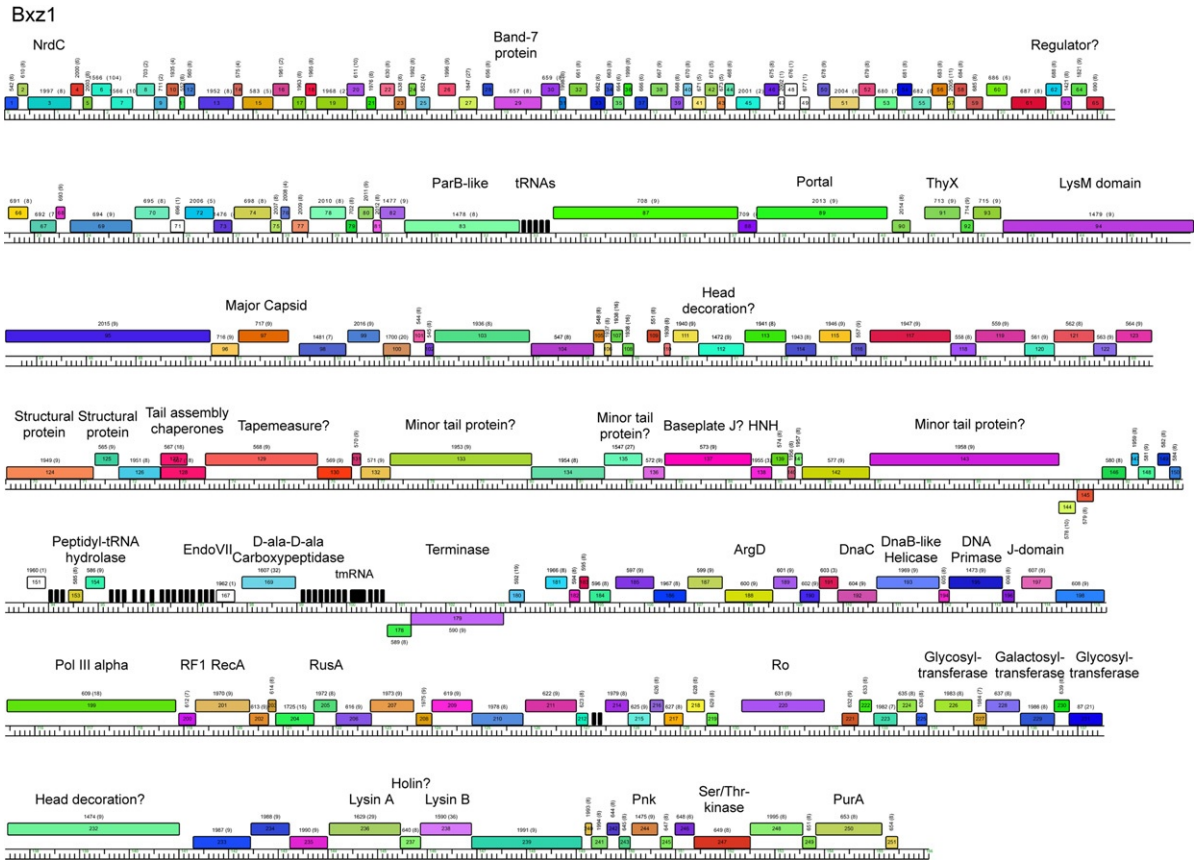


FIGURE 6 Map of the phage Bxz1 genome, a member of Subcluster C1. See Figure 3A for further details on genome map presentation.

Author's personal copy

none of the Cluster C homologues share this intein, although relatives are present in some terminase genes of nonmycobacteriophage phages, as well as host DnaB-like helicases. The third ET08 intein is in gp202, which is also present in homologues in ScottMcG, Spud, Catera, and Rizal, as well as in some Cluster A terminases (see Section III.A). ET08 gp239 contains an intein present in all Cluster C homologues except for Bxz1 gp239 and Myrna 246, but is related (28% amino acid identity) to the fifth ET08 intein, which is in gp248; none of the Cluster C homologues of ET08 gp248 contain this intein. Because Cluster C phages are not apparently temperate, it is no surprise that they do not encode integrase or partitioning functions. However, it is striking that one of these—phage LRRHood—carries a close relative of the Cluster A repressors (Pope *et al.*, 2011). Specifically, LRRHood gp44 is near identical to the known Bxb1 repressor (gp69) and differs in only one amino acid substitution (see Section III.A, Section V.A.1, and Fig. 20). However, the LRRHood genome does not contain even a single copy of the 13-bp stoperator sequence, which is the known binding site for this repressor; this, therefore, does not appear to be an immunity system for LRRHood. Presumably LRRHood has ''stolen'' the repressor from a Cluster A1-like phage, and because there are only between 5- and 11-bp differences between LRRHood gene *44* and Cluster A1 repressor genes, this presumably occurred relatively recently in evolutionary time. A plausible role for LRRHood gp44 might be to exclude other phages from superinfecting cells that are undergoing lytic growth of the phage. We note that another example of apparent repressor theft is seen in Cluster F phages (see Sections III.F and V.A.1).

Although the Subcluster C2 phage Myrna shares its myoviral morphology and genome size with Subcluster C1 phages, it is substantially different at the genomic level. There are segments where synteny is maintained, including regions corresponding to Bxz1 genes *117–143* and *187–211*, but other regions, including the leftmost ~31 kbp, have very few genes in common. These Cluster C phages therefore represent a really intriguing collection of viruses whose genomes suggest many secrets waiting to be discovered.

D. Cluster D

There are six closely related phages in Cluster D, and no subcluster divisions (Table I). These phages form plaques that are not completely clear, and although not evidently turbid as phages such as L5 or Bxb1, it is possible that they are temperate and form lysogens at low frequency. However, the genomes reveal no features associated with temperate phages such as integrases or repressors. The genomes are presumed to be circularly permuted and terminally redundant, and the left end is arbitrarily designated at a noncoding gap 6-7 genes to the left of the

Author's personal copy

terminase large subunit gene; a gene map of the Cluster D representative, PBI1, is shown in Figure 7. All of the PBI1 genes, except for *41* and *42*, are transcribed rightward (Fig. 7), and the virion structure and assembly genes are located in the left part of the genome (*8-31*); although confident assignments can be ascribed to the terminase large subunit gene (*8*), portal (*9*), tapemeasure (*27*), and minor tail proteins (*30*, *31*), assignments of capsid subunit (*17*) and major tail subunit genes (*24*) are less confident. None of the virion proteins have been characterized experimentally. The lysis cassette, including lysin A and lysin B genes (*35* and *39*, respectively), is located immediately downstream of the virion structure and assembly operon (Fig. 7). PBI1 genes *36-38* all encode proteins with putative transmembrane domains (two, four, and one, respectively) and at least one may act as a holin. None of the Cluster D phages infect *M. tuberculosis*. Genes in the right half of the genome contain several that are likely involved in DNA replication, including a helicase (*50*), a DNA polymerase III α subunit (*63*), and a primase/polymerase gene (*81*). Several other genes in this region are implicated in nucleotide metabolism, including a putative nucleotide-binding protein (*56*), a putative deoxyribonucleotidase (*59*), and ThyX (*65*). Gene *57* encodes a 309 residue protein containing a highly acidic segment in its C-terminal half (76 of 79 resides are aspartic or glutamic acids) and has weak similarity to CDC45-related proteins, implicating it in a possible role in initiation of DNA replication.

A particularly intriguing aspect of Cluster D phages is genes immediately to the left of the terminase large subunit gene (*8*). Five of the six Cluster D phages (PBI1, Adjutor, P-lot, Butterscotch, and Troll4) each encode a 256 residue protein (PBI gp7; Fig. 7) with similarity to vegetative insecticidal protein 2 (VIP2) family proteins encoding actin-ADP-ribosylating toxins (Han *et al.*, 1999). In phage Gumball, the gene encoding VIP2 is part of a single gene (*6*) that also contains sequences corresponding to PBI1 gene *6*, as a result of an additional A-residue immediately upstream of the VIP2 moiety. Although this mutation could have arisen during propagation of the phage in the laboratory and perhaps renders it nonfunctional, an alternative possibility is that PBI1 gene *6* encodes a function related to VIP2-like activity and that the two units can function either as independent proteins (as in PBI gp6 and gp7) or as two domains of a single protein (as in Gumball gp6). None of these proteins contain putative signal sequences and it is plausible that they are expressed late in lytic growth (together with their closely linked virion structure and assembly genes) and are released upon cell lysis, a parallel scenario to the expression and release of toxin from *E. coli* phage 933W (Tyler *et al.*, 2004). In light of the similarity to insecticidal toxins, it is puzzling as to what role is played by PBI1 gp7 and its related proteins. Do Cluster D phages confer insecticidal properties to infected cells and, if so, what is the natural bacterial host and what insects are affected? These questions remain unresolved.

Author's personal copy

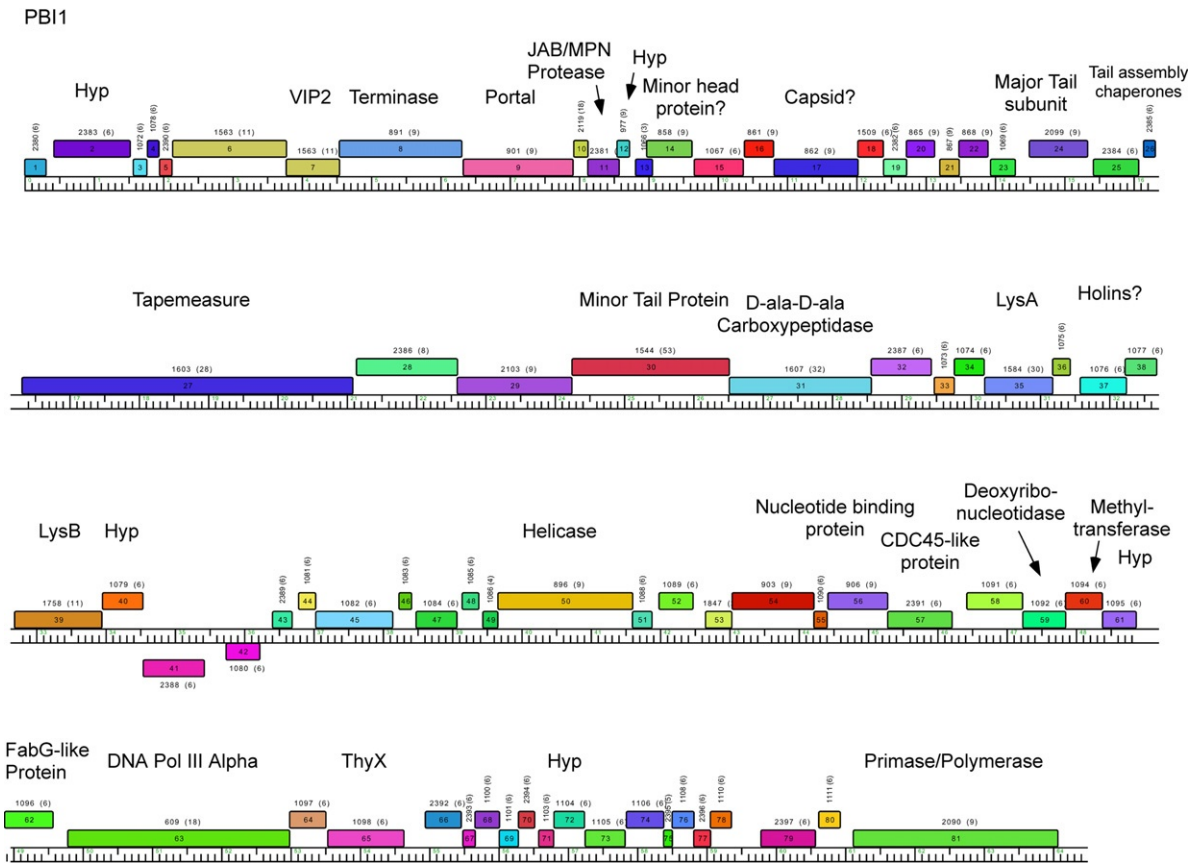


FIGURE 7 Map of the phage PBI1 genome, a member of Cluster D. See Figure 3A for further details on genome map presentation.

Author's personal copy

E. Cluster E

There are five Cluster E phages, all are very similar to each other at the nucleotide sequence level—albeit with a variety of replacements, insertions, and deletions—and there are no subcluster divisions (Table I). They form plaques with some turbidity and likely form lysogens, although stable lysogens have not been reported for any Cluster E phages. Preliminary data from our laboratory suggest that Cjw1 may form stable lysogens at 42°C, but not at 37°C. None of these phages infect *M. tuberculosis*. A map of the genome organization of the Cluster E representative phage, Cjw1, is shown in Figure 8. It encodes both small and large subunits of terminase, and the small subunit gene is approximately 1 kbp away from the physical left end of the genome (Fig. 9). The terminase large subunit genes of Cjw1 and Kostya (gp8 and gp9 respectively) both contain inteins related to those in some of the Cluster A1 terminases (see Section III.A). The virion structure and assembly operon spans from genes *5* to *28* and follows the canonical synteny, but with two notable interruptions. First, in Pumpkin, Porky, and 244, but not in Cjw1 or Kostya, an EndoVII Holiday junction resolvase gene lies between the frameshifting tail assembly chaperones and the tapemeasure protein. Although this is an unusual position for a gene insertion, presence of a HJ resolvase in the virion structural operon is reminiscent of the organization of Cluster B genomes (see Section III.B). Presumably it is not essential for growth because Cjw1 and Kostya lack this gene and no obvious candidates exist for providing this function elsewhere in their genomes. Second, in Cjw1 (and Pumpkin and 244), two small genes are located between the terminase small and large subunit genes, transcribed in the opposite direction and of unknown function (Fig. 8); one of these (Cjw1 *17*) is absent from Kostya and Porky. Genes related to Cjw1 *8* are also found in Cluster B4 genomes (e.g., Cooper *98*), elevating confidence in their assignment. The lysis cassette lies to the right of the virion structure and assembly genes and includes both Lysin A and Lysin B genes (Cjw1 *32* and *35*, respectively); Cjw1 gene *33* is a strong candidate for encoding a holin with two membrane-spanning domains. Genes to the right of this appear to be organized into six putative operons: (1) leftward-transcribed genes *38–52*, (2) genes *53* (integrase) and *54* that are transcribed rightward, (3) leftward-transcribed genes *55* and *56*, (4) genes *57–128* transcribed rightward, (5) leftward-transcribed genes *129–142*, and (6) genes *143* and *144* that are transcribed rightward. In general, this organization is conserved in all Cluster E phages. Although the vast majority of genes in Cluster E genomes are of unknown function (>80%), several of those that do have functional assignments are rare among the mycobacteriophages and are of considerable interest. First, Cjw1 gene *39* encodes a relative of Lsr2, a regulatory protein present in the mycobacterial host that coordinates expression of a

Author's personal copy

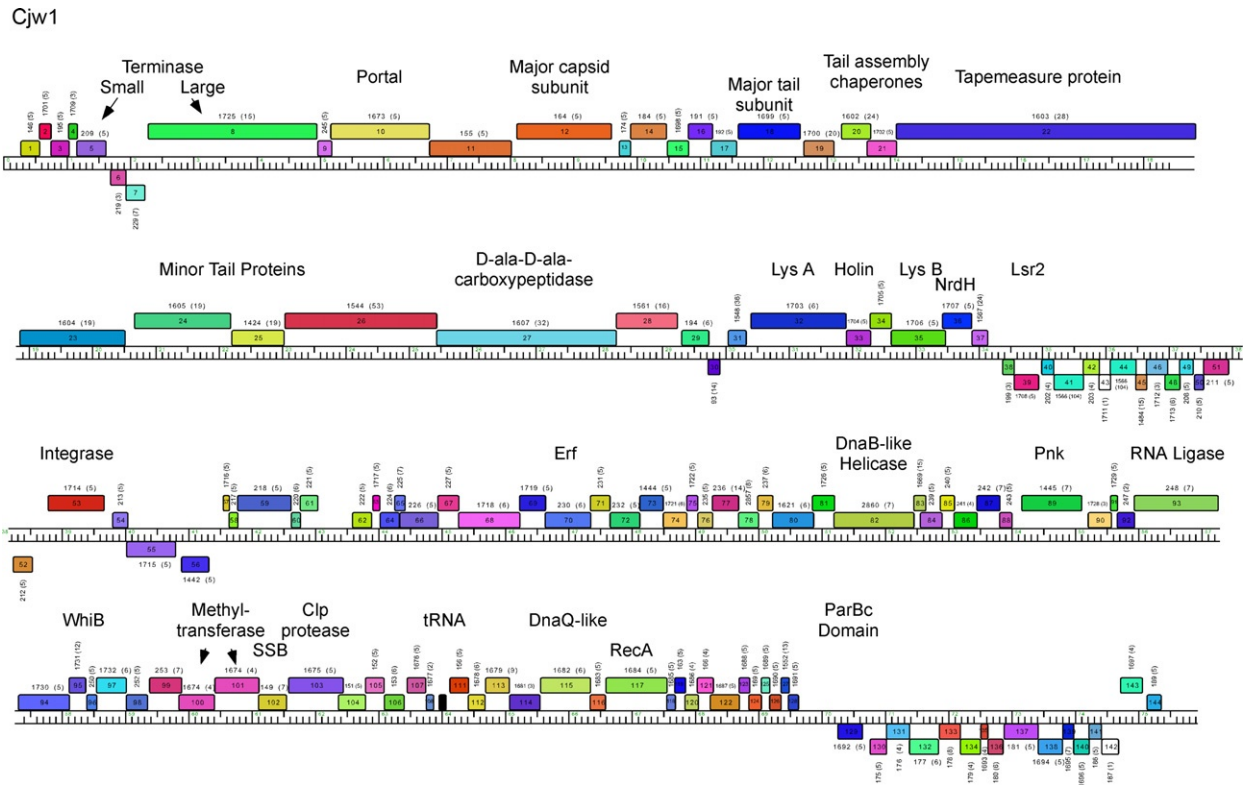


FIGURE 8 Map of the phage Cjw1 genome, a member of Cluster E. See Figure 3A for further details on genome map presentation.

Author's personal copy

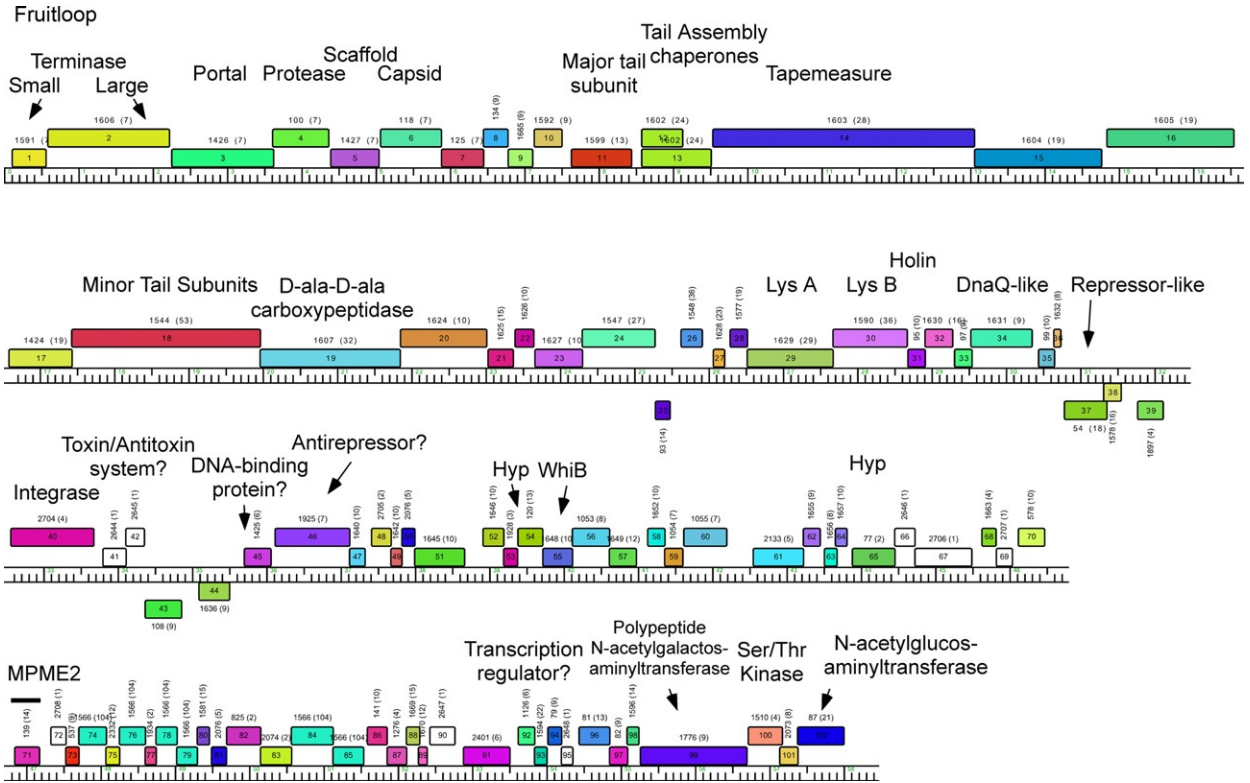


FIGURE 9 Map of the phage Fruitloop genome, a member of Subcluster F1. See Figure 3A for further details on genome map presentation.

Author's personal copy

large number of host genes. Its role in Cjw1 is not clear, although we note that Cluster J genomes also encode related proteins (see Sections III.J). It is plausible that this confers the repressor function in Cluster E phages, although this would be a very unusual form of phage regulation.

Second, the long operon *57–128* includes several genes encoding putative functions in DNA replication, recombination, RNA metabolism, and nucleotide metabolism. Cjw1 gene *70* is related to Erf-family recombinases and presumably mediates general recombination functions. Although several other mycobacteriophages encode RecA-like and RecET-like recombination functions, Cluster E phages—along with Omega (Cluster L) and Wildcat—are the only ones with an Erf-like protein; somewhat surprisingly, Cluster E phages also encode RecA homologues (e.g., Cjw1 gp117). Cjw1 encodes a $tRNA^{gly}$ gene (*109*) in this region, as well as a tRNA-like gene in the small intergenic gap between genes *108* and *109*, although the noncanonical tRNA structure and four-base anticodon suggest that this is either nonfunctional or perhaps plays a role in translational frameshifting. The RNA Ligase encoded by Cjw1 gene *93* is also unusual but related proteins are also encoded in Cluster L and J phages. Likewise, Cjw1 gene *102* encoding a single-stranded binding protein (SSB) is only found elsewhere in Cluster L phages and the singleton Wildcat. Cjw1 gp89 has been shown to be a bifunctional polynucleotide kinase (Pnk) with both kinase and phosphatase domains, and it was noted that because Cjw1—like Omega (see Section III.J)—also encodes an RNA Ligase (gp93), that these might act to evade an RNA-damaging antiviral host response (Zhu *et al.*, 2004). Interestingly, we note that the Cluster L phage LeBron also encodes similar Pnk and RNA Ligase proteins (see Section III.L; Fig. 15). Moreover, this highly diverse set of phages also encodes one or more tRNA genes (Table I). Finally, Cjw1 gene *115* encodes a protein with similarity to DnaQ-like proteins, suggesting a possible role in DNA repair or perhaps in phage replication itself. Roles for these interesting proteins, their biochemical activities, and their expression patterns await elucidation.

F. Cluster F

Cluster F is one of the more diverse groups of phages at the nucleotide sequence level. There are a total of 10 members, with all but 1 (Che9d) constituting Subcluster F1, and Che9d forming Subcluster F2. They form somewhat turbid plaques from which stable lysogens can be recovered (Pham *et al.*, 2007). Genomes vary somewhat in length, ranging from 52.1 kbp [Ardmore, (Henry *et al.*, 2010b)] to 59.5 kbp [Che8 (Pedulla *et al.*, 2003)](Table I), but all have defined cohesive termini. None of the Cluster F phages infect *M. tuberculosis*. The complete sequence of mycobacteriophage Ms6 is not yet available, but from sequenced segments of

Author's personal copy

the genome it seems probable that it belongs to the F cluster. The genome map of the Cluster F representative phage, Fruitloop, is shown in Figure 9. Fruitloop encodes both terminase small and large subunits, and the small subunit gene is very close to the physical left end of the genome (Fig. 9). The virion structure and assembly operon extends from gene *1* to gene *24*, transcribed rightward, and is fairly canonical with regard to the common syntenic organization (Fig. 9). The block of genes corresponding to the region from Fruitloop *11* (major ail subunit) to gene *23* (putative minor tail protein) is the most highly conversed segment among Cluster F1 phages at the nucleotide level. The region to the left of Fruitloop gene *11* is substantially different in Ramsey and Boomer, although the genes are likely to confer similar functions in DNA packaging and head assembly. The lysis cassette lies to the right of the virion structure and assembly genes and includes lysin A (*29*) and lysin B (*30*) genes; gp31 is a likely Holin and contains a single predicted transmembrane domain (Fig 9). Immediately to the right of the lysis cassette is a DnaQ-like gene (*34*) implicated either in DNA repair, or perhaps DNA replication itself, as in Cluster E phages.

Genes in the right part of the Fruitloop genome are organized into four possible operons: (1) genes *37–39*, (2) genes *40–42*, (3) genes *43* and *44*, and (4) genes *45–102*. The first of these is of particular interest, as Fruitloop genes *37–39* do not have closely related counterparts in other Cluster F phages, but are homologues to genes in Cluster A phages; gp37 and gp38 are close relatives of Bxb1 gp69 and gp70, respectively; and gp39 is most closely related to Jasper gp92. Bxb1 gp69 is a well-characterized repressor related to the L5 repressor (Jain and Hatfull, 2000) and its presence in Fruitloop is somewhat surprising (see Sections III.A and V.A.1). Two lines of evidence suggest that Fruitloop gp37 is not involved directly in the immunity regulation of Fruitloop itself, in that it is absent from all other Cluster F genomes, and there is not an abundant array of stoperator sites throughout the Fruitloop genome as there are in Bxb1 and its relatives (see Section III.A). There is, however, a single putative 13-bp repressor-binding site located upstream of gene *39* near a strongly predicted putative leftward promoter, which thus may be involved in autoregulation of its expression. We also note that the nucleotide sequences of Fruitloop *37* and *38* are ~98% identical with their Bxb1 homologues, suggesting that these were acquired very recently in evolutionary times. A plausible scenario is that Fruitloop has stolen these genes from a Cluster A-like phage for purposes of conferring a rogue immunity status, providing protection to Fruitloop lysogens from superinfection by Cluster A-type phages that have a Bxb1 type of immunity. A similar example of apparent repressor theft occurs in the Subcluster C1 phage LRRHood (see Section III.C).

The rightward operon encompassing genes *40–42* contains the integrase gene (*40*) plus two genes (*41*, *42*) that are candidates for forming a toxin–antitoxin (TA) system; Fruitloop gp41 is the putative toxin and gp42

Author's personal copy

is the putative antitoxin. There are no identifiable relatives of these in other Cluster F phages or indeed in any other mycobacteriophages. TA systems generally are not common in phage genomes, although the well-studied plasmid addiction system of phage P1 is within this general class (Lehnherr *et al.*, 1993). However, it seems unlikely that genes *41* and *42* are involved in plasmid maintenance of Fruitloop similar to P1 because Fruitloop encodes an integrase (gp40) and presumably provides prophage maintenance through stable integration. Because it has been reported that TA systems can provide protection to bacterial cultures by conferring abortive infection (Fineran *et al.*, 2009), an intriguing hypothesis is that this Fruitloop TA system has been acquired to provide protection to Fruitloop lysogens by infection from other phages. In this model, addition of the Bxb1 repressor and the putative TA system has been selected for by the same core property of providing survival of the host to subsequent viral attack.

The vast majority of genes in the Fruitloop rightward operon containing genes *45–102* are of unknown function, but several are of interest. First, gene *45* encodes a helix-turn-helix DNA-binding protein, which could either provide repressor activity or possibly a Cro-like function. Second, Fruitloop gene *55* encodes a WhiB-family transcriptional regulator protein, and although WhiB-related proteins are encoded by several mycobacteriophages, their roles remain unclear (Rybniker *et al.*, 2010). Third, gene *100* encodes a putative serine-threonine kinase of unknown function, and it is unclear whether it is phosphorylating host or phage proteins. Fourth, there are two putative genes encoding glycosyltransferase enzymes, although the roles and the targets of these are also unknown. Finally, gene *71* is part of a MycobacterioPhage Mobile Element (MPME2)(see Sections III.G and IV) that is prevalent throughout Cluster F phages and was first identified through genome comparison of Cluster G phages (Sampson *et al.*, 2009).

G. Cluster G

Four Cluster G phages are extremely closely related to each other at the nucleotide sequence level and there are no subcluster divisions. These phages have among the smallest mycobacteriophage genomes, ranging from 41.1 to 42.3 kbp in length. As discussed later, the primary cause for length differences is the presence/absence of a novel small putative mobile genetic element (MPME), which is absent from Angel and present as a single copy but in a different location in each of the other three genomes. Cluster G phages form lightly turbid plaques from which stable lysogens can be recovered (Sampson *et al.*, 2009). They do not infect *M. tuberculosis* at high efficiency, but mutants arise at a frequency of $\sim 10^{-5}$ that have acquired the ability to infect *M. tuberculosis* at equal efficiency to

Author's personal copy

M. smegmatis (Sampson *et al.*, 2009). A genome map of the Cluster G representative phage, BPs, is shown in Figure 10. Cluster G genomes have defined cohesive ends, and a putative small terminase subunit gene (*1*) is located near the left physical end of the genome (Fig. 10). The virion structure and assembly operon encompasses genes *1* through *26* and is organized with canonical synteny. The lysis cassette follows immediately after and contains both lysin A and lysin B genes (*27* and *28*, respectively; Fig.10); gene *29* encodes a putative holin, based on the presence of two putative transmembrane domains. The only leftward-transcribed genes in the genome are *32* and *33*, encoding the integrase and repressor proteins, respectively, and genes to their right are all transcribed rightward. Most of the genes in this rightward operon are of unknown function, although it also includes a RecET recombination system (gp42 and gp43) and a RuvC-like Holliday Junction resolvase (gp51).

A rather striking feature of repressor/integrase gene organization is that the crossover site for integrase-mediated, site-specific recombination within the phage attachment site (*attP*) is located within the coding region for the repressor (Sampson *et al.*, 2009). As a consequence, two different types of gene product are expressed from gene *33*: a 130 residue product from the viral genome and a 97 residue product from an integrated prophage. The 97 residue protein confers immunity and provides the repressor function, whereas the virally expressed 130 residue protein does not. Integration and excision would therefore seem to play critical roles in the decision between lysogenic and lytic growth, having a direct impact on whether an active or an inactive repressor is expressed (Sampson *et al.*, 2009). A particularly interesting question arises as to how the genetic switch operates and whether phage-encoded cII and/or cIII analogues modulate the frequencies of lysogeny or whether this is accomplished solely by the gene *32–33* cassette (see Section V.A.2).

The close nucleotide sequence similarity between Cluster G genomes proved crucial in identification of a new class of ultrasmall mobile genetic elements present in mycobacteriophages (Sampson *et al.*, 2009). For example, when BP is compared with the other three genomes, it is apparent that the small open reading frames BPs *57* and *59* form a single gene in Angel, Hope, and Halo (Pope *et al.*, 2011; Sampson *et al.*, 2009). Alignment of DNA sequences shows that there is a precise insertion of 445 bp, including the open reading frame for gene *58*, in BPs relative to the other phages (Figs. 10B and 10C). Such alignments also show similar relationships reflecting an insertion in Halo that has occurred at a target within the homologue of BP gene *54* and in Hope at a target within the homologue of BP gene *56* (Fig. 10C). Alignment of the inserted sequences shows that there are two types of these MPME elements, MPME1 (in Hope and BPs) and MPME2 (in Halo), that share 78% nucleotide sequence

Author's personal copy

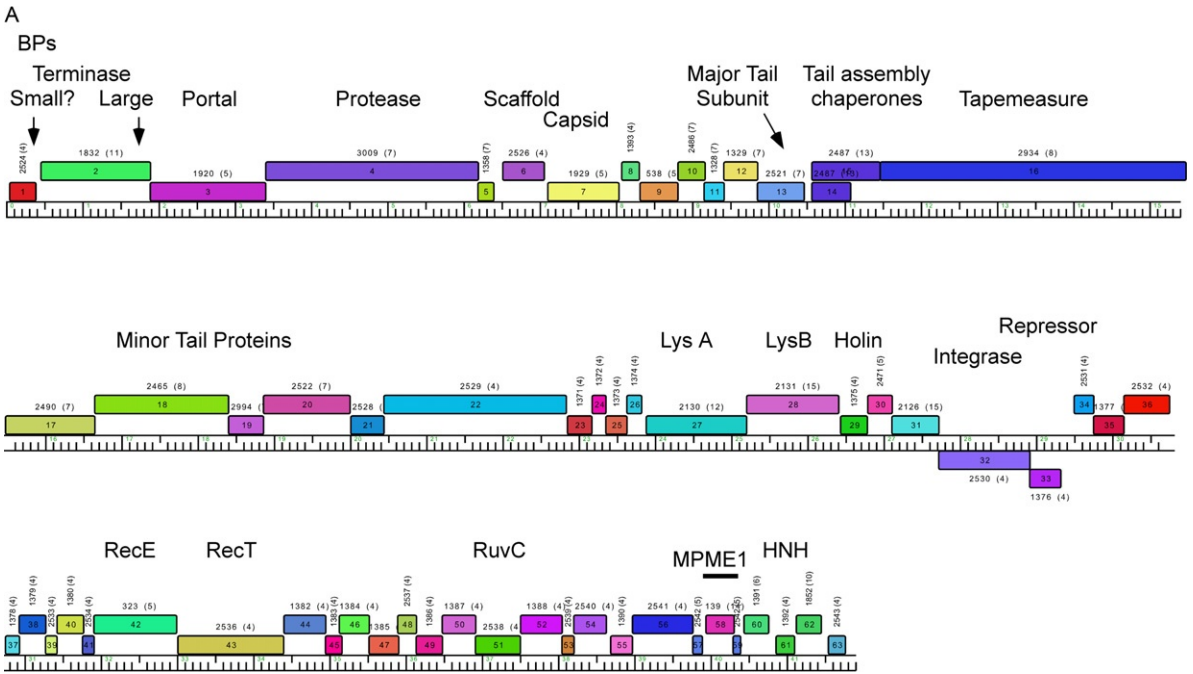


FIGURE 10 (Continued)

Author's personal copy

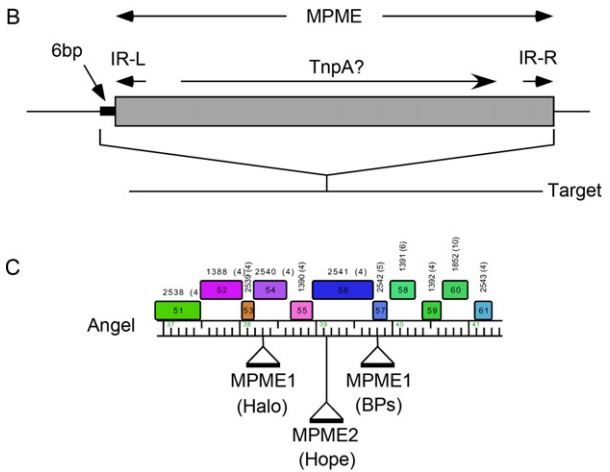


FIGURE 10 Features of Cluster G phages. (A) Map of the phage BPs genome, a member of Cluster G. See Figure 3A for further details on genome map presentation. (B) MPME elements are 439–440 bp in length and are flanked by 11-bp imperfect inverted repeats, IR-L and IR-R. Insertion into the target is associated by a 6-bp insertion between IR-L and the target, and its origin is unknown. (C) A 5.5-kbp segment at the extreme right end of the Angel genome is shown, illustrating the positions of the insertion of MPME1 elements in BPs and Halo, and an MPME2 insertion in phage Hope.

Author's personal copy

similarity (Sampson *et al.*, 2009); Angel is devoid of these elements (Sampson *et al.*, 2009). Comparison against other mycobacteriophages shows that there is a single copy of MPME1 in Cluster F phages Fruitloop, PMC, Llij, Boomer, Che8, Tweety, Ardmore and Pacc40; in Cluster I phages Brujita and Island 3; and a partial copy in Corndog. There are no related copies in any of the sequenced mycobacterial genomes or elsewhere (Sampson *et al.*, 2009).

Although the size of each of the MPME1 insertions is 445 bp, the mobile element itself appears to be 439 bp long, with two imperfect 11-bp inserted repeats (IR) at the extreme ends (Sampson *et al.*, 2009) (Fig. 10B). At the right end, IR-R is joined to the target sequence without addition or duplication of any target sequences (Fig.10B). However, at the left junction, there is an insertion of 6 bp between IR-L and the target DNA. This 6-bp segment is different in many of the insertions and does not correspond to target duplication. It therefore remains a mystery as to where this 6 bp originates from and what mechanism of transposition could be involved in generating these types of products. Transposition is presumably mediated by the 123 residue product encoded within the MPME element (Fig. 10B), although this is both remarkably small and shows no motifs to structural elements common to other transposases.

The MPME elements show that the three genes containing insertions (corresponding to BPs genes *54*, *56*, and the gene *57–59* interruption; Fig. 10) are presumably nonessential for phage growth. However, because there are no obvious differences in growth of the four Cluster G phages, this provides little information as to what the genes actually do and why they may have been acquired by the phages—they are simply nonessential. Cluster G genomes are suitable substrates for BRED manipulation (see Section VII.A.6), and four additional genes (BPs genes *44*, *49*, *50*, and *52*) have also been shown to be nonessential because viable deletion mutants can be constructed readily. This raises the question as to whether it is generally true that a high proportion of genes constituting the nonvirion structure and assembly genes are nonessential in the mycobacteriophages and, if so, what forces drive the evolutionary of this large number of mysterious genes.

H. Cluster H

There are three phages assigned to Cluster H: two (Predator and Konstantine) in Subcluster H1 and one (Barnyard) in Subcluster H2 (Hatfull *et al.*, 2010; Pope *et al.*, 2011). The cluster is quite diverse and many differences exist among the three constituting genomes. These phages form plaques that are not evidently turbid, but also not completely clear, although stable lysogens have not been recovered; the genomes also do not possess features of temperate phages such as integrase or

Author's personal copy

repressor genes. They all have termini consistent with the viral chromosomes being circularly permuted and terminally redundant. They have among the lowest of the GC% content of the mycobacteriophages (Fig. 2), and none of them infect *M. tuberculosis*.

The genomic organization of a Cluster H representative, Barnyard, is shown in Figure 11. The left end of Barnyard is designated arbitrarily at the first noncoding interval to the left of the terminase large subunit gene (*6*), and the functions of the five intervening genes are unknown; none of these is an obvious candidate for a terminase small subunit gene (Fig. 11). All of the predicted genes are transcribed in the rightward direction as shown in Figure 11, and an obvious genomic feature is the presence of the large number of orphams (genes belonging to a phamily that has only a single member), nearly 60% of all 109 predicted Barnyard genes. Such a high proportion of orphams is not unexpected in singleton phage genomes where other closely related phages have yet to be isolated, but for Barnyard this reflects the degree to which it—as a Subcluster H2 phage—differs from Subcluster H1 phages. Subcluster H1 phages Predator and Konstantine have 18 and 25% orphams, respectively, again reflecting the generally high diversity of this cluster. This is in marked contrast to, for example, Cluster G phages, which differ mostly by just a relatively modest number of nucleotide differences (see Section III.G).

The virion structure and assembly genes span from gene *6* through to gene *36*, and genes encoding the terminase large subunit (*6*), portal (*7*), a putative protease (*17*), capsid subunit (*21*), major tail subunit (*30*), tail assembly chaperones expressed by a putative programmed translational frameshift (*31* and *32*), tapemeasure protein, and putative minor tail proteins (*34-36*) are predicted (Fig. 11). Although these genes are in canonical order, several small additional genes are present between the putative portal and protease genes, and also between the capsid and major tail subunit genes. The tapemeasure gene is notable due to its impressive length (6.1 kbp), corresponding to the very long tails of the Cluster H phages (~300 nm). In Predator—although not the other Cluster H phages—there is a putative Endo VII Holliday Junction resolvase (*10*) located between the portal and protease genes, reminiscent of the location of functionally related genes in some Cluster B and E genomes. The lysis cassette, containing lysin A (*39*) and lysin B (*40*) genes, as well as a putative holin (*41*), lies immediately to the right of the virion structure and assembly operon (Fig. 11). Of the 67 genes in the Barnyard genome to the right of the lysis cassette, only 13 have homologues in other mycobacteriophages; 8 of these are found only in Subcluster H1 phages. Five genes in this region can be assigned putative functions, including a Helicase (*65*), a putative nucleotide-binding protein (*75*), an α subunit of DNA polymerase III (*80*), a peptidase (*94*), and a large primase/polymerase gene (*108*).

Author's personal copy

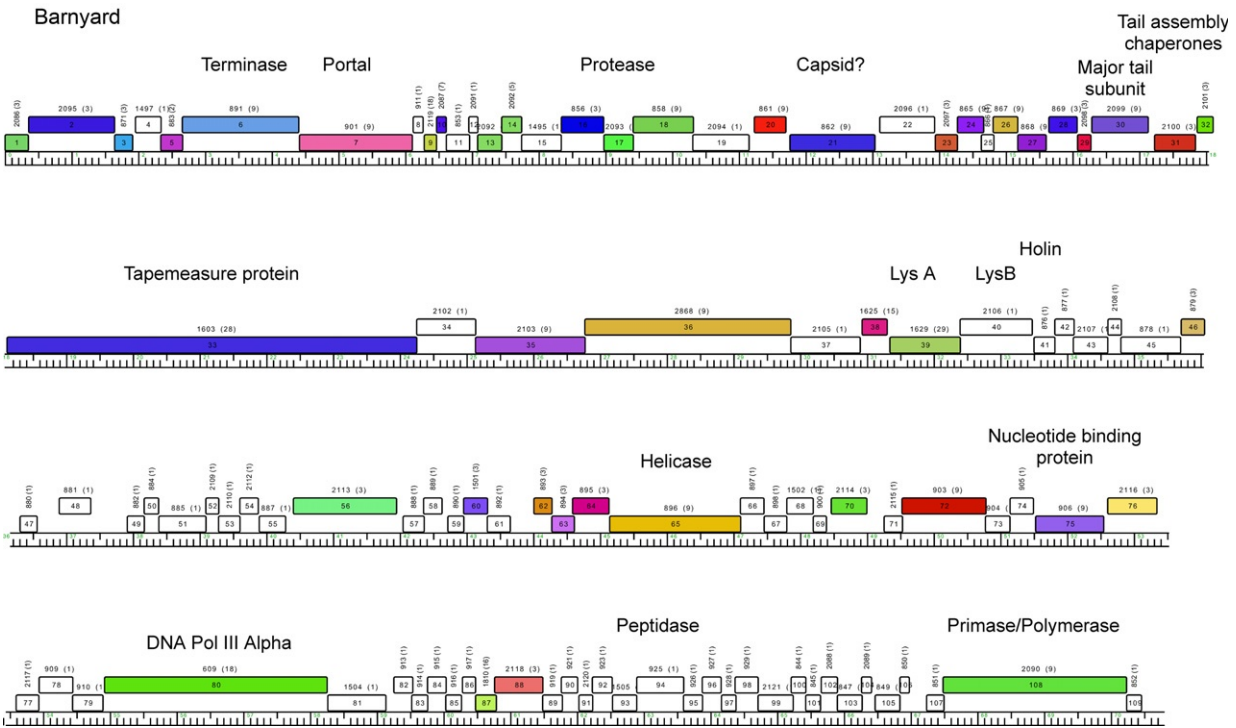


FIGURE 11 Map of the phage Barnyard genome, a member of Subcluster H1. See Figure 3A for further details on genome map presentation.

Author's personal copy

The GC% of Cluster H phages is among the lowest of all the mycobacteriophages (56.3–57.3%, Fig. 2), and only the singleton Wildcat shares a GC% content lower than 60% (56.9%). This may reflect a preference of the Cluster H phages for hosts that are more distantly related to *M. smegmatis*, and although the majority of members of the Actinomycetales have GC% contents that are above 60%, some—such as *M. leprae* (57.8%)—do have a substantially lower GC% content. Such a different host preference may account for notable differences of Cluster H phages from other mycobacteriophages and the high proportions of orphams (Fig. 11). Cluster H phages thus remain largely unexplored and would seem to warrant extensive further analysis both in regards to the determination of gene function and expression and in elucidation of their host ranges.

I. Cluster I

Cluster I contains three phage members: two (Brujita and Island3) in Subcluster I1 and one (Che9c) in Subcluster I2. They are quite diverse at the sequence level, although the two Subcluster I1 phages share nucleotide sequence similarity across most of their genomes. Che9c is both more distantly related and has a substantially larger genome (57 kbp) than Subcluster I1 genomes (~47 kbp). Cluster I phages form somewhat turbid plaques, although lysogens have not been well characterized. They do, however, encode genes common to temperate phages such as an integration system; no repressor genes have been described. All Cluster I phages have defined cohesive termini, and gene *1* is a reasonable candidate for encoding a terminase small subunit (Fig. 12). None of the Cluster I phages infect *M. tuberculosis*. A notable morphological feature of Cluster I phages is that they contain prolate heads, with a length to width ratio of approximately 2.5:1 (Hatfull *et al.*, 2010). The genome organization of a Cluster I representative, Che9c, is shown in Figure 12. The virion structure and assembly operon (gene *1–22*) is syntenically canonical, and genes encoding terminase large subunit (*2*), portal (*4*), protease (*5*), capsid (*6*), major tail subunit (*12*), tail assembly chaperones (*13* and *14*), tapemeasure (*15*), and minor tail proteins (*16–19*) can be predicted confidently (Fig. 12). The lysis cassette lies to the right of the virion structure and assembly operon and includes genes encoding lysin A (*25*), lysin B (*26*), and a putative holin (27) containing two predicted membrane-spanning domains. To the right of the lysis genes are perhaps four operons: (1) gene *30* and *31* transcribed leftward, (2) genes *32–36* transcribed rightward, (3) genes *37–46* transcribed leftward (although two large intergenic regions exist between genes *39* and *40* and between *45* and *56* so there could be multiple operons; Fig. 12), and (4) genes *47–84* transcribed rightward.

Author's personal copy

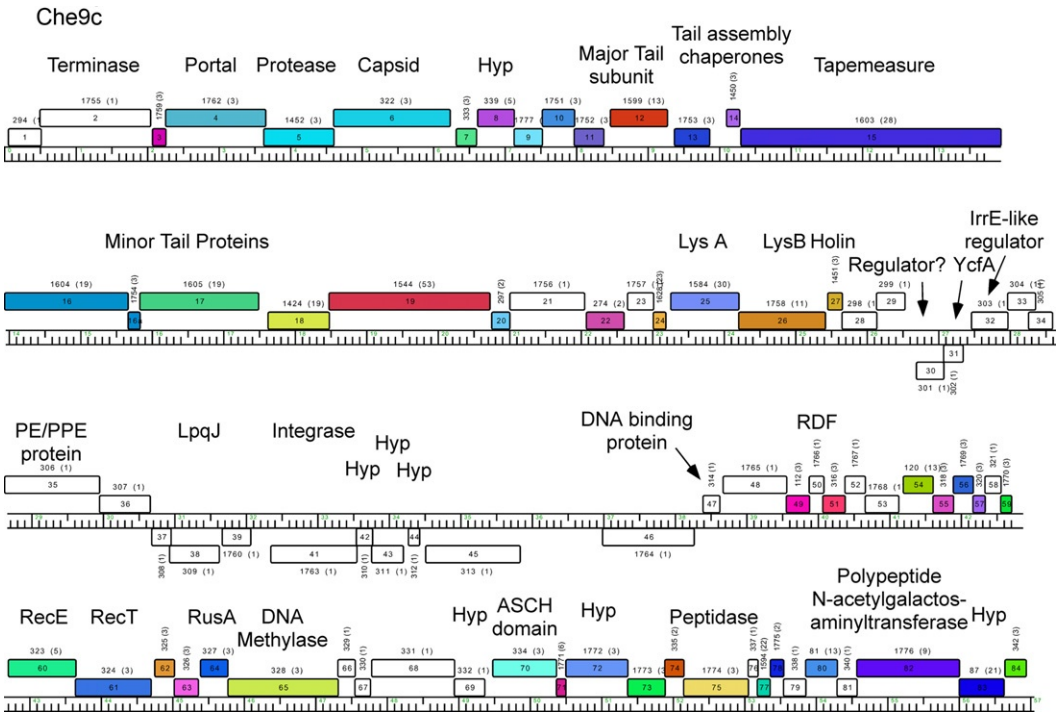


FIGURE 12 Map of the phage Che9c genome, a member of Subcluster I2. See Figure 3A for further details on genome map presentation.

Author's personal copy

Several of the genes in these operons encode putative transcriptional regulators, including genes *30*, *32*, *46*, and *47*. Che9c gp32 is unusual in that it is related to the IrrE regulator of *Deinococcus radiodurans* and there are no similar genes elsewhere in the mycobacteriophages; it contains both putative DNA recognition and protease motifs. Gene *46* encodes a large protein with a putative helix-turn-helix motif near its N terminus, and gene *47* encodes a smaller helix-turn-helix containing a predicted DNA-binding protein. Any of these could plausibly play the role of the phage repressor, but it is curious that there are such a variety of putative regulatory proteins. Che9c gene *41* encodes a tyrosine-integrase and a putative Xis encoded by gene *50*, displaced almost 7 kbp from the *int* gene. A putative *attP* common core can be identified immediately adjacent to the integrase gene, and Che9c is predicted to integrate at an *attB* site overlapping a host $tRNA^{tyr}$ gene (see Section V.B.1 and Table II). Cluster I1 phages have a different integration specificity and are predicted to integrate into a $tRNA^{thr}$ gene (see Section V.B.1 and Table II). The *attP* site of Che9c is notable because whereas other phages that integrate into tRNA genes carry the 3′ end of the host tRNA, Che9c is predicted to encode a complete $tRNA^{tyr}$ gene at this position. However, the predicted tRNA has a number of nonstandard features that bring into question whether this is either expressed or functional, and it is possible that it is just a bioinformatic quirk.

The segment to the left of the Che9c integrase contains several genes of interest, including one encoding a putative PE/PPE-like protein (*35*), and an LpqJ-like predicted lipoprotein containing a single transmembrane domain; gp38 is also a predicted membrane protein with three membrane-spanning domains. Although nothing is known about the expression patterns of Che9c or any other Cluster I phages, it is tempting to suggest that these genes were acquired relatively recently from a bacterial chromosome through an errant excision process and are expressed from an integrated prophage, perhaps conferring new properties to lysogenic strains. Che9c gp38 is related more closely to the LpqJ protein of *M. smegmatis* (Msmeg_0704 product) than to other bacteria, but these share only 42% amino acid sequence identity, making it unlikely that it was a recent acquisition from this host specifically. The region to the right of integrase contains five leftward-transcribed genes, all of which are orphams, and genes *42*, *43*, and *44* are all related to large families of hypothetical bacterial proteins of unknown function; there is an unusually large noncoding region (~1.1 kbp) between genes *45* and *46*. The GC % content of the 4.7-kbp gene *42–46* region is substantially different (59.9%) from the overall GC% of the Che9c genome (65.4%), consistent with the interpretation that it has been acquired relatively recently by horizontal genetic exchange, most likely from a bacterial host. Furthermore, the genome organization of Che9c is similar to Subcluster I1 phages

Author's personal copy

to the left of Che9c gene *19*, and the similarity—although still rather weak—does not pick up again until to the right of Che9c gene *51*. Differences in lengths of the intervening regions account for the 10-kbp differences in overall genome lengths. The rightward-transcribed operon containing genes *47–64* encodes a number of genes of interest. These include RecET-like genes (*60* and *61*), which are of note because they have been exploited to develop a system for recombineering in mycobacteria (van Kessel and Hatfull, 2007, 2008a,b; van Kessel *et al*., 2008) and of the mycobacteriophages themselves (Marinelli *et al*., 2008)(see Section VII.A.6). Gene *64* encodes a RusA-like Holliday Junction resolvase, gene *75* encodes a putative peptidase, and gene *82* encodes a protein predicted to encode polypeptide *N*-acetylgalactosaminyltransferase activity; Subcluster F1 and the Subcluster C2 phage Myrna encode similar enzymes, and it is of considerable interest to identify which proteins—either phage or perhaps host encoded—are targets of glycosylation. Interestingly, Subcluster I1 phages Brujita and Island3 lack homologues of Che9c gene *82*, but at the right ends of their genomes encode a protein with a different sequence but which is predicted to have the same activity as Che9c gp82. Cluster I genomes clearly are rich in features of interest and warrant substantial further investigation.

J. Cluster J

Cluster J contains the published genome Omega and unpublished phages LittleE and Baka; the genome organization of Omega is shown in Figure 13. Omega forms slightly turbid plaques from which stable lysogens can be recovered (G. Broussard and Graham F. Hatfull, unpublished results) and does not infect *M. tuberculosis*. The genome is 110 kbp long and contains defined cohesive termini, although with unusually short 4-base single-stranded DNA extensions (Pedulla *et al*., 2003)(see Section VI.A). The left end is ~1.5 kbp from the putative terminase small subunit gene, and the virion structure and assembly genes extend to approximately gene *44* (Fig. 13). Genes encoding terminase small (*3*) and large subunits (*11*), portal (*13*), protease (*14*), capsid (*15*), major tail subunit (*31*), tail assembly chaperones expressed via a programmed translational frameshift (*32* and *33*), tapemeasure (*34*), and minor tail proteins (*35–40*) can be predicted with reasonable confidence (Fig. 13); the terminase large subunit contains an intein similar to that in some Cluster A and E terminases (see Sections III.A and III.E). However, this operon contains many interruptions with insertions of genes transcribed both in forward and reverse directions (Fig. 13). For example, there are seven small open reading frames (*4–10*) of unknown function between terminase small and large subunit genes, and immediately to the right of capsid genes are open reading frames encoding putative glycosyltransferase and *O*-methyltransferase activities

Author's personal copy

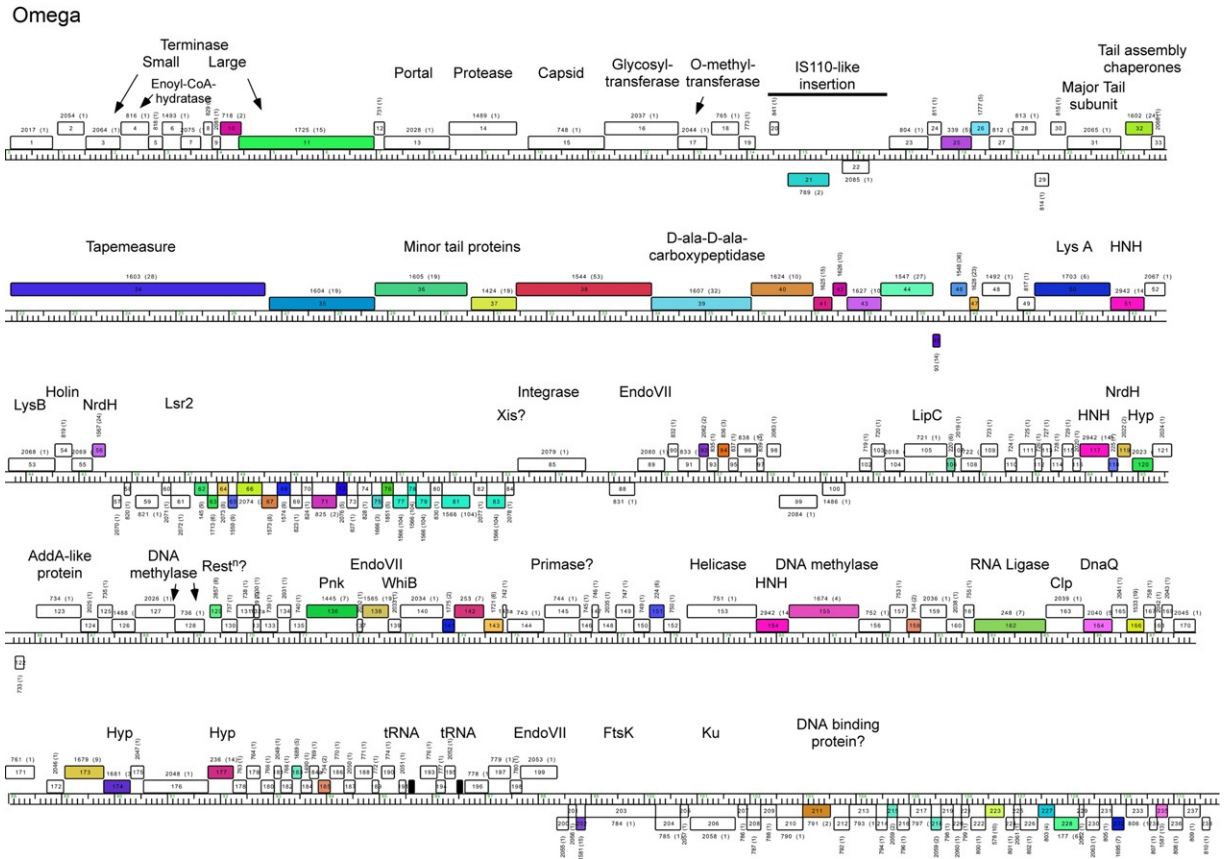


FIGURE 13 Map of the phage Omega genome, a member of Cluster J. See Figure 3A for further details on genome map presentation.

Author's personal copy

(Fig. 13). It is unclear what the specific functions of these genes are, although their location within the virion structure and assembly operon suggests the intriguing possibility that they are modifying virion proteins. Leftward-transcribed genes *21* and *22* are of unknown function, although there are homologues of both in the Subcluster A1 phage, Bethlehem (gp71 and gp72)(see Section III.A). Omega gp21 also has weak sequence similarity to IS110 family transposases, and thus both Omega genes *21* and *22* could conceivably belong to an uncharacterized IS110-like transposon. The absence of this segment in the LittleE genome—as well as the related insertion in Bethlehem—strongly supports this possibility.

The Omega lysis cassette lies to the right of the virion structure and assembly genes, and includes lysin A (*50*) and lysin B (*53*) genes—separated by an HNH domain gene (*51*) and a gene of unknown function (*52*)—as well as a putative holin (*54*). To the right of this is a leftward-transcribed operon containing 28 small open reading frames. Accurate annotation of the many small genes in phage genomes is an ongoing challenge, but this operon presents a good example of the utility of comparative genomic analyses because more than half of these are related to genes in mycobacteriophages in other clusters and subclusters (Fig. 13); Omega genes *77*, *79*, *81*, and *83* all belong to a relatively large phamily with representatives of the total of 104 members in virtually all clusters of phages. Gene *61* is of interest because it encodes a homologue of the host *lsr2* gene, which has been shown to be a global regulator of gene expression in *M. tuberculosis* (Colangeli *et al.*, 2007, 2009)(see Section III.E). It is unclear what functional role there could be for Omega gp16, perhaps acting as a regulator of phage gene expression, but more enticingly as a possible regulator that reprograms host gene expression of lysogenic strains.

Although the Omega genome is replete with orphams and genes of unknown function, several genes with predicted functions make them odd denizens of a phage genome. For example, gene *206* encodes a homologue of bacterial Ku-like proteins involved in mediating nonhomologous end joining (NHEJ)(Pitcher *et al.*, 2007). Ku-like proteins typically act together with a dedicated DNA ligase (Lig IV), which is absent from the Omega genome (Pitcher *et al.*, 2006). Interestingly, the NHEJ system seems to be required for Omega infection and presumably is required for genome recircularization upon infection (Pitcher *et al.*, 2006). Further details are provided in Section VI.B. We note that the only other mycobacteriophage to encode a Ku-like protein is the singleton mycobacteriophage Corndog (gp87)(see Section III.M.1) and it too has 4-base ssDNA extensions (Pitcher *et al.*, 2006).

Just to the left of the Ku-like protein gene, gene *203* encodes an FtsK-like protein. Bacterial FtsK proteins, including that of *M. smegmatis*, contain three domains: an N-terminal domain involved in membrane association; a central domain containing motor functions, including a

Author's personal copy

AAA-ATPase motif; and a short C-terminal domain (gamma) that confers the specificity of DNA binding (Sivanathan *et al.*, 2006). Its primary function in bacteria is to facilitate proper segregation of daughter chromosomes at cell division. Omega gp203 is a 443 residue protein that lacks the N-terminal domain of bacterial FtsK proteins and contains just the core domain and a putative C-terminal gamma domain. However, although the core domain is quite closely related to that of *M. smegmatis* FtsK (60% amino acid identity), the gamma domain is distinctly different and is not closely related to the gamma domains of any other known FtsK proteins. The presence of this FtsK-like gene in a phage genome is highly unusual, and its role is unknown (Pedulla *et al.*, 2003). It is possible that it acts on a host chromosome that contains different 8-bp asymmetric FtsK Orienting Polar Sequences (KOPS) targeting sequences for gp203 recognition, and although such a host has apparently yet to be described genomically, it is not obviously *M. smegmatis*. Alternatively, it could be acting on the Omega genome itself, and it would be of interest to determine bioinformatically or experimentally if Omega contains KOPS-like gp203-binding sites. We note, however, that gp203 is unlikely to simply facilitate partitioning of extrachromosomally replicating prophage molecules because Omega encodes an integrase (gp85), as well as a putative excise (gp84), and Omega lysogens contain an integrated prophage (G. Broussard and Graham F. Hatfull, unpublished results)(Fig. 13). Omega also encodes a number of proteins predicted to be involved in DNA metabolism, including a DnaQ-like protein (*183*), DNA methylases (*127*, *128*, *165*), and an AddA-like protein. Curiously, it encodes three proteins with sequence similarity to EndoVII Holliday Junction resolvases (*89*, *138*, and *199*). It is unclear why any phage genome would need to encode HJ resolvase activity in three separate genes. Omega—with its curious collection of genes with predicted functions and its vast array of hundreds of genes of unknown function—clearly warrants much more detailed investigations to understand gene expression and gene function and how these contribute to the overall biology of this phage and its Cluster J relatives. Omega also encodes a bifunctional polynucleotide kinase (gp136, Pnk) similar to that of Cjw1 and is proposed to act with the RNA Ligase (gp162) to evade an RNA-damaging host response (Zhu *et al.*, 2004). Omega encodes two putative tRNAs: a $tRNA^{gly}$ (gene *192*) closely related to the one in Cjw1 and a noncanonical putative tRNA with a 4-base anticodon also similar to that encoded by Cjw1 (see Section III.E).

K. Cluster K

Cluster K contains three genomes divided into two subclusters: K1 and K2. Subcluster K1 contains Angelica and CrimD, and Subcluster K2 contains TM4; three additional unpublished phages also belong to this cluster

Author's personal copy

(Pope *et al.*, 2011). TM4 and its derivatives are perhaps the most widely utilized in mycobacterial genetics, and a map of TM4 genome organization is shown in Figure 14. All of the Cluster K phages infect *M. tuberculosis* as well as *M. smegmatis*, and TM4 was originally isolated by recovery from a putative lysogenic strain of *Mycobacterium avium* (Timme and Brennan, 1984). TM4 forms clear plaques on mycobacterial lawns, whereas the other Cluster K phages form turbid plaques from which stable lysogens can be recovered.

Cluster K phages contain defined cohesive termini (Table I), and the terminase large subunit gene is located near the physical left end (Fig. 14). All of the genes are transcribed in the rightward direction, with the exception of genes *39–41*. Genes *4* through *25* encode the virion structure and assembly functions, and genes encoding terminase large subunit (*4*), portal (*5*), protease (*6*), scaffold (*8*), capsid (*9*), major tail subunit (*14*), tail assembly chaperones expressed via a programmed translational frameshift (*15*, *16*), tapemeasure protein (*17*), and minor tail proteins (*18–25*) can be predicted with reasonable confidence (Fig. 14). To the right of the structural genes lies the lysis cassette, and this included both lysin A (*29*) and lysin B (*30*) genes, as well as a gene (*31*) encoding a putative Holin that has four predicted transmembrane domains. The remainder of the genome encodes several genes with predicted functions that are of interest. TM4 gene *49* encodes a putative WhiB-like protein, and it is noteworthy that WhiB family proteins are found in a variety of the mycobacteriophages, including phages of Clusters E, F, and J (see Sections III.E, III.F, and III.J). Ryniker and colleagues (2010) showed that TM4 gp49 is highly expressed soon after infection and functions as a dominant negative regulator of the host WhiB2 protein, and when TM4 gp49 is expressed in *M. smegmatis* it induces septation inhibition. TM4 gene *69* is not an essential gene for viral propagation (Rybniker *et al.*, 2010) but is implicated in mediating superinfection exclusion.

TM4 also encodes a large putative Primase/Helicase enzyme (gp70) and a RusA-like Holliday Junction resolvase (gp71). TM4 gene *79* encodes an SprT-like protease of unknown function, although it is noteworthy that a number of mycobacteriophages encode proteases of a variety of types in the nonstructural parts of their genomes; phages in Clusters A, C, and K also encode SprT-like proteases, Cluster E phages encode a Clp-like protease, and Cluster I phages encode a predicted peptidase. These are distinct from proteases encoded as part of the virion structural gene operon where they play a role in capsid assembly, although a variety of different types of enzymes appear to perform that function. Presumably there are protease-required processing events involved outside of capsid assembly, but these remain poorly understood. At least 3 to 4 kbp of the TM4 genome must be nonessential for growth because a variety of shuttle phasmids have been constructed in which parts of the genome are

Author's personal copy

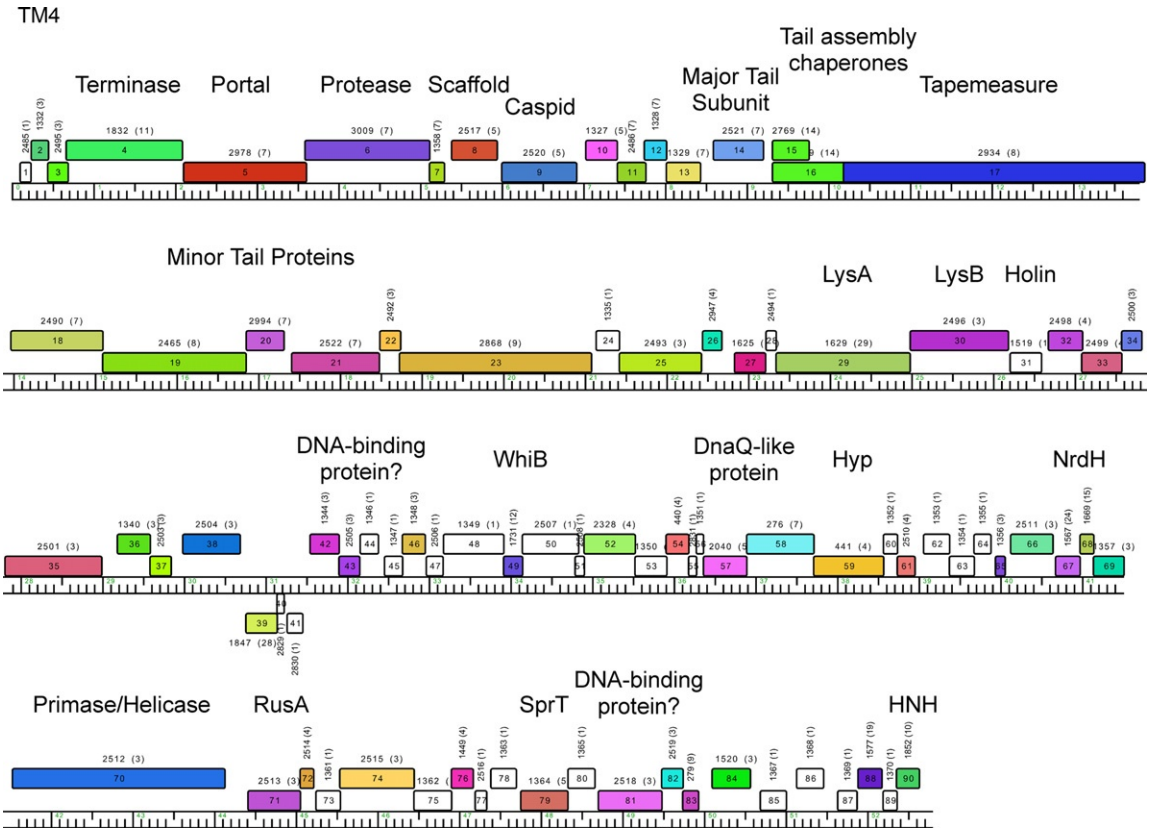


FIGURE 14 Map of the phage TM4 genome, a member of Subcluster K2. See Figure 3A for further details on genome map presentation.

Author's personal copy

replaced by a cosmid vector (Bardarov *et al*., 1997; Jacobs *et al*., 1987, 1989). The extent of the deleted regions is not yet clear but is within the right half of the genome.

The two Subcluster K1 phages, Angelica and CrimD, are quite similar to each other and both are rather different from TM4 at the nucleotide sequence level (Pope *et al*., 2011). Nonetheless, many of the genes are homologues when compared at the amino acid level, especially in the virion structure and assembly operons. A notable difference though is that Cluster K1 genomes are approximately 7 kbp larger than TM4 (Table I); this difference is largely accounted for by a large insertion in the middle of K1 genomes relative to TM4. It thus seems likely that TM4 acquired a large central deletion, possibly during its time of isolation, a scenario reminiscent of the properties of some of the Cluster A phages, such as D29 (see Section III.A). Unfortunately, K1 and K2 genomes are insufficiently similar at the nucleotide level to determine precisely how such a deletion might have occurred. The central segment of K1 phage genomes encodes an integrase and a plausible transcriptional regulator, consistent with the idea that the nontemperate behavior of TM4 arises as a consequence of this deletion. Putative *attP* sites are located adjacent to their integrase genes, and Angelica and CrimD are predicted to integrate at an *attB* site overlapping the host tmRNA gene (see Section V.B.1 and Table II). These are the only mycobacteriophages known to use this integration site (*attB-9*, Table II) and all others that encode a tyrosine-integrase integrate into known host tRNA genes (see Section V.B.1 and Table II). We also note that both Angelica and CrimD encode a $tRNA^{trp}$ gene (gene *5*) located between the terminase large subunit gene (*8*) and the left physical end. This tRNA gene is similar to $tRNA^{trp}$ genes encoded by L5 and D29 (95% identity across 59 of the 75 bp; see Section III.A) as well as the *M. smegmatis* mc^2155 host tRNA gene (Msmeg_1343, 90% over 73 bp) and presumably could have been acquired from either a phage or a host genome. Derivatives of TM4 are perhaps the most widely used mycobacteriophages in mycobacterial genetics because of their ability to infect both fast- and slow-growing mycobacteria and the availability of TM4 shuttle phasmids that can be manipulated readily (Jacobs *et al*., 1987). Shuttle phasmids are chimeras containing a mycobacteriophage moiety and an *E. coli* cosmid moiety, such that they can be propagated as large plasmids in *E. coli* and as phages in mycobacteria (Jacobs *et al*., 1987, 1991). Construction of shuttle phasmids involves a step in which these chimeras are packaged into phage λ heads *in vitro* (Jacobs *et al*., 1991) and thus it is not surprising that TM4 shuttle phasmids contain deletions of phage DNA such as to accommodate the cosmid vector insertion. TM4 shuttle phasmids can be manipulated readily using standard genetic and molecular biology approaches in *E. coli* and have been exploited for the delivery of transposons (Bardarov *et al*., 1997), reporter

Author's personal copy

genes (Jacobs *et al.*, 1993), and allelic exchange substrates (Bardarov *et al.*, 2002)(see Section VII.A).

L. Cluster L

Cluster L contains just a single published phage genome, LeBron; however, six additional unpublished phages also fall within Cluster L. LeBron is anticipated to be competent to form lysogens and encodes its own tyrosine-integrase (Pope *et al.*, 2011). It is relatively recently isolated and little is known about its general biological properties. A map of the LeBron genome organization is shown in Figure 15.

A LeBron gene encoding a terminase large subunit (*4*) is located approximately 1.0 kbp from the physical left end of the genome, and there are three small open reading frames predicted in the intervening region. One of these (*2*) is a candidate for encoding a terminase small subunit, and has sequence similarity to a gene immediately upstream of the terminase large subunit gene in Omega (Fig. 13). Within the putative virion structure and assembly operon (*2–24*) genes encoding a terminase large subunit (*4*), portal (*5*), protease (*6*), capsid (*7*), major tail subunit (*13*), tail assembly chaperones (*14* and *15*), tapemeasure protein, and minor tail proteins (*17–24*) are predicted. The lysis cassette follows this and includes both lysin A (*25*) and lysin B (*26*) genes and the putative holin gene *27*. To the right there are short leftward-transcribed operons, including a tyrosine-integrase gene, followed by a longer rightward-transcribed operon that includes several genes implicated in regulation and nucleotide or DNA metabolism. These include a putative WhiB regulator (gp78), a kinase (gp74), a ssDNA-binding protein (gp81), an RNA ligase (gp65), a putative DNA Pol II (gp93), EndoVII (gp96) and RusA (gp76) Holliday Junction resolvases, a DnaB-like helicase, a ribosephosphate kinase (gp71), and an Erf-like general recombinase (gp60). The kinase may act with the RNA Ligase to evade an RNA-damaging host response as proposed to Cjw1 and Omega (Zhu *et al.*, 2004). Two additional genes encode a putative esterase (gp46) of unknown specificity and a nicotinate phosphoribosyltransferase (gp72). The rightmost 10 kbp of the LeBron genome contains mostly leftward-transcribed genes of unknown function, with the exception of gene *128* that encodes an AAA-ATPase protein. About 80% of LeBron genes remain of unknown function. LeBron also encodes nine tRNA genes ($tRNA^{leu}$, $tRNA^{thr}$, $tRNA^{lys}$, $tRNA^{tyr}$, $tRNA^{trp}$, $tRNA^{leu}$, $tRNA^{his}$, $tRNA^{cys}$, and $tRNA^{lys}$) of unknown function, but we note that while most of these correspond to codons used highly in the LeBron genome (see Section III.A), one of them ($tRNA^{leu}$) has an anticodon corresponding to the rare codon 5′-CUA. Overall, LeBron is an interesting genome with numerous genes, suggesting an intriguing but poorly understood biology.

Author's personal copy

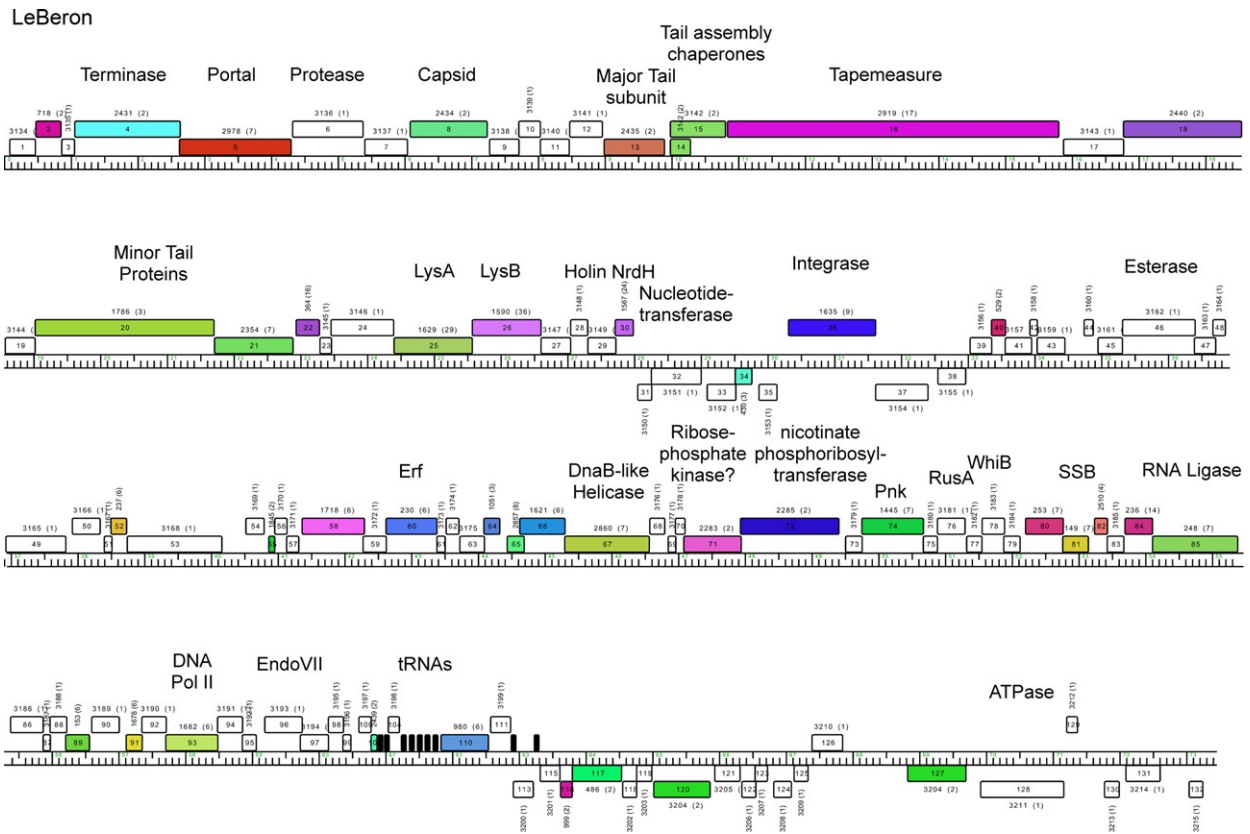


FIGURE 15 Map of the phage LeBron genome, a member of Cluster L. See Figure 3A for further details on genome map presentation.

Author's personal copy

M. Singletons

1. Corndog

The singleton phage Corndog has an unusual morphology with a prolate head having approximately a 4:1 length-to-width ratio. It looks like a corndog! Corndog forms plaques on *M. smegmatis* that are neither completely clear nor turbid, and stable lysogens have not been reported. Corndog does not infect *M. tuberculosis*. A map of the Corndog genome organization is shown in Figure 16.

Corndog contains a 69.8-kbp genome with defined cohesive ends, having 4-base ssDNA extensions as described earlier for Omega (see Section III.J). However, the putative terminase large subunit gene (*32*) is located ~13.5 kbp away and there are 31 predicted genes between it and the left physical end. Most of these 31 genes are of unknown function and the majority are orphams. However, genes *6* and *7* have regions associated with DNA methylases, gene *11* encodes an Endo VII Holiday Junction resolvase, gene *22* encodes a primase/polymerase, and gene *29* has an HNH domain. Gene *25* is part of a truncated copy of an MPME1 mobile element (Sampson *et al.*, 2009)(see Fig. 10B). Genes *1–12* are organized into an apparent leftward-transcribed operon, whereas genes *13–31* are transcribed rightward and could be part of the viral structure and assembly operon that continues to the right.

The Corndog virion structure and assembly genes (*32–67*) containing genes encoding the terminase large subunit (*32*), portal (*34*), protease (*39*), capsid (*41*), major tail subunit (*49*), tail assembly chaperones expressed via a programmed translational frameshift (*54* and *55*), tapemeasure protein (*57*), and minor tail proteins (*58-67*) can be predicted confidently. Although these genes appear in the canonical order, their synteny is disrupted in at least five locations. Two of these involve HNH insertions (genes *33* and *56*) and one (*40*) is a single small gene inserted between the putative protease and capsid genes that does not appear similar to scaffold proteins (Fig. 16). A fourth is between the putative major tail subunit gene and the tail assembly chaperones and contains four small open reading frames, one of which (*51*) has homologues in some Cluster C phages. Another of these (*53*) is a predicted transcriptional regulator. The fifth syntenic interruption is between the putative portal protein and the protease (Fig. 16). Four open reading frames are present (genes *35–38*), and three of them (*35*, *37*, and *38*) are predicted to encode an *O*-methytransferase, a polypeptide *N*-acetylgalactose aminyltransferase, and a glycosyltransferase, respectively. These are similar functions to genes located within the virion structure and assembly operon of Omega (Fig. 13), except that in Omega they are inserted between the capsid and major tail subunit genes. It is not known if these enzymes are responsible for modification of virions, although this is an intriguing possibility that warrants investigation.

Author's personal copy

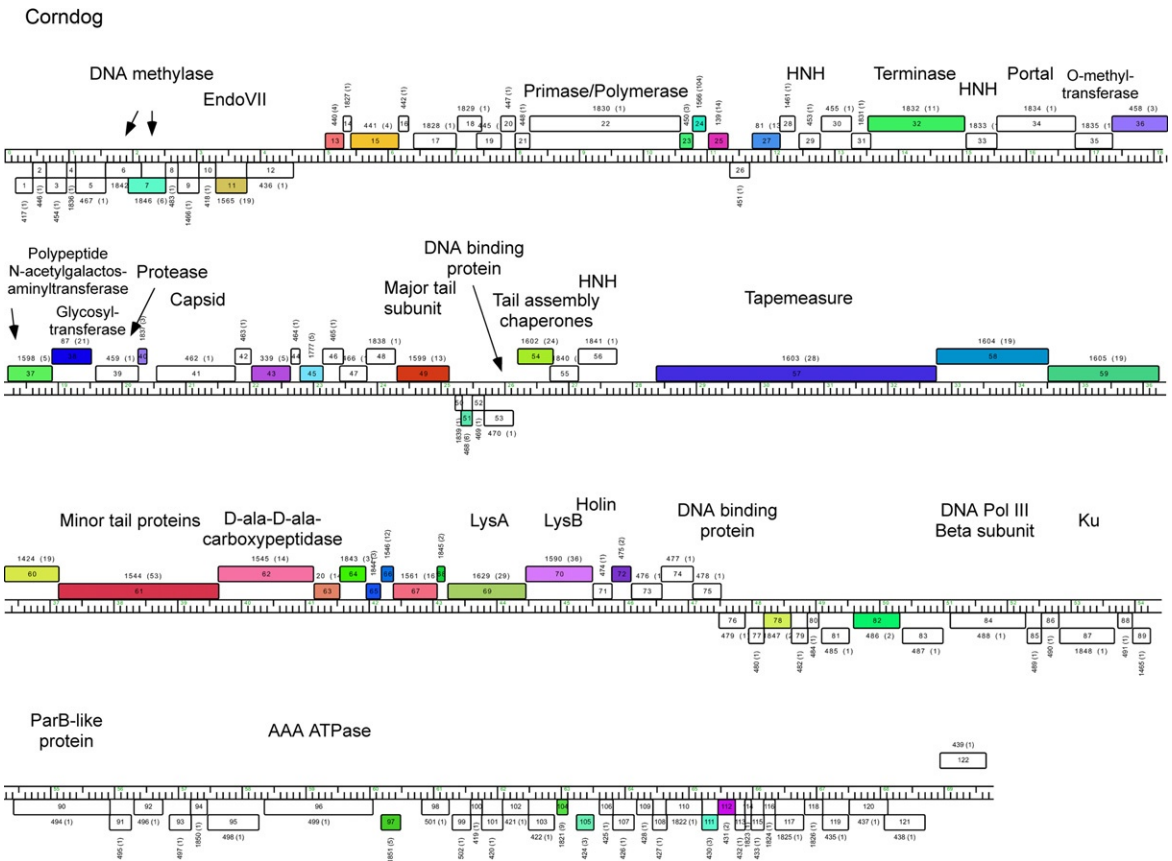


FIGURE 16 Map of the singleton phage Corndog genome. See Figure 3A for further details on genome map presentation.

Author's personal copy

Sequences of the virion structure genes provide few clues to the unusual Corndog prolate head morphology. The putative capsid subunit (gp41) is not closely related to any other mycobacteriophage-encoded proteins, and its closest homologue is a gene within the *Bifidobacterium dentium* genome, although the two proteins are only 26% identical. It does, however, contain a predicted capsid domain and HHPred reports similarity to the HK97 capsid structure. We note that the Corndog capsid subunit sequence has no evident sequence similarity with the capsid subunits of Cluster I phages, which also have prolate heads but with a different length:width ratio (see Section III.I).

The lysis cassette lies to the right of the virion structure and assembly operon and contains lysin A (*69*) and lysin B (*70*) genes, as well as a putative holin (*71*). To the right of that is a long leftward-transcribed operon (*76–121*) amazingly enriched for orphams (only 7 of the 57 genes have evidently related genes in other mycobacteriophages), although several genes in this operon have interesting predicted functions (Fig. 16). First, Corndog gene *87* encodes a Ku-like protein distantly related to both the *M. smegmatis* homologue (Msmeg_5580; 35% amino acid sequence identity) and the Ku-like gp206 protein encoded in phage Omega (32% amino acid sequence identity) (Pitcher *et al.*, 2006). Both Corndog and Omega share the features of having cohesive genome termini with 4-base extensions and encoding Ku-like proteins, supporting the idea that Ku-like proteins play a role in NHEJ-mediated genome circularization upon infection (see Sections III.J and VI.B). Second, Corndog gene *82* encodes a putative DNA polymerase III β subunit implicated in acting as a loading clamp in DNA replication. It is unclear whether this gene is required for Corndog replication and what role it could play. However, this is the only occurrence of this particular function in any of the mycobacteriophage genomes. It is not obvious that it was acquired recently from a bacterial host, in that its closest relative is the clamp loader of *Saccharopolyspora erythraea* and the proteins share only 23% amino acid sequence identity. There are no other closely related phage-encoded homologues. A third gene of interest is Corndog *96* encoding an AAA-ATPase, although it is not closely related to other AAA-ATPases encoded by other mycobacteriophages (such as LeBron gp128, Myrna gp262, or Che8 gp69). Its origins are also unclear, and the closest homologue is a protein encoded by *Haliangium ochraceum*, which shares 33% amino acid sequence identity. There are no closely related homologues in other phage genomes.

Finally, Corndog gp90 is a ParB-like protein implicated in chromosome partitioning. A plausible role for such a function could be to provide stability to an extrachromosomally replicating Corndog prophage, although because lysogens have not yet been recovered, this is unclear. An alternative possibility is that this protein plays a different role such as

Author's personal copy

a regulatory function rather than partitioning per se. We note in this regard that there is no putative ParA protein, a striking difference from the putative partitioning functions in RedRock (see Fig. 3B and Section III.A). There is no obvious homologue of Corndog gp90 in *M. smegmatis*, but there are related genes in other mycobacterial strains and other Actinomycetales genomes, with the closest homologue being *Mycobacterium kansasii* Spo0J (46% amino acid sequence identity).

2. Giles

The singleton Giles forms lightly turbid plaques on *M. smegmatis* from which stable lysogens can be recovered (Morris *et al.*, 2008). Giles does not infect *M. tuberculosis*. A map showing Giles genome organization is shown in Figure 17. The genome has defined cohesive termini with long (14-base) single-stranded DNA extensions. Like other singleton genomes, Giles contains a high proportion of orphams and only 14 of the predicted 78 genes have readily identifiable homologues in other mycobacteriophages (Fig. 17). Giles gene *1* corresponds to a possible terminase small subunit gene, but is separated by three short open reading frames on the opposite strand from the terminase large subunit gene (Fig. 17). The putative virion structure and assembly operon extends from gene *1* to gene *36* and has several striking features. First, most of these are orphams, reflecting the considerable sequence divergence from other mycobacteriophages. Genes encoding the terminase large subunit (*5*), portal (*6*), protease (*7*), capsid (*9*), tail assembly chaperones expressed via a programmed frameshift (*17* and *18*), tapemeasure (*19*), and minor tail proteins (*21–28*, *36*) can be identified readily (Fig. 17). Gene *8* may encode a scaffold-like protein, based on its position in the operon, but no major tail subunit gene can be predicted confidently. Second, the lysis cassette lies upstream of gene *36*, which has been shown experimentally to be a virion-associated protein, and thus the lysis cassette—including lysin A (*31*), lysin B (*32*), and putative holin genes (*33*)—appears to lie within this operon (Morris *et al.*, 2008). In most other mycobacteriophages genomes (with the notable exception of Cluster A phages), it is noteworthy that the location of the lysis cassette is immediately downstream of the virion structure and assembly operon, and because the virion proteins have been characterized experimentally for few of these phages, it is plausible that they may also have genes downstream of the lysis cassette that encode virion proteins; it is also plausible that Giles gene *36* was acquired relatively recently, and we note that other mycobacteriophage genomes contain tail genes in noncanonical positions [e.g., L5 gene *6*; (Hatfull and Sarkis, 1993)]. A more curious feature of this part of the Giles genome is the presence of the integration cassette—including *integrase* and *xis* genes, as well as *attP*—between the minor tail subunit genes and the lysis cassette (Morris *et al.*, 2008). It is plausible that this cassette relocated to

Author's personal copy

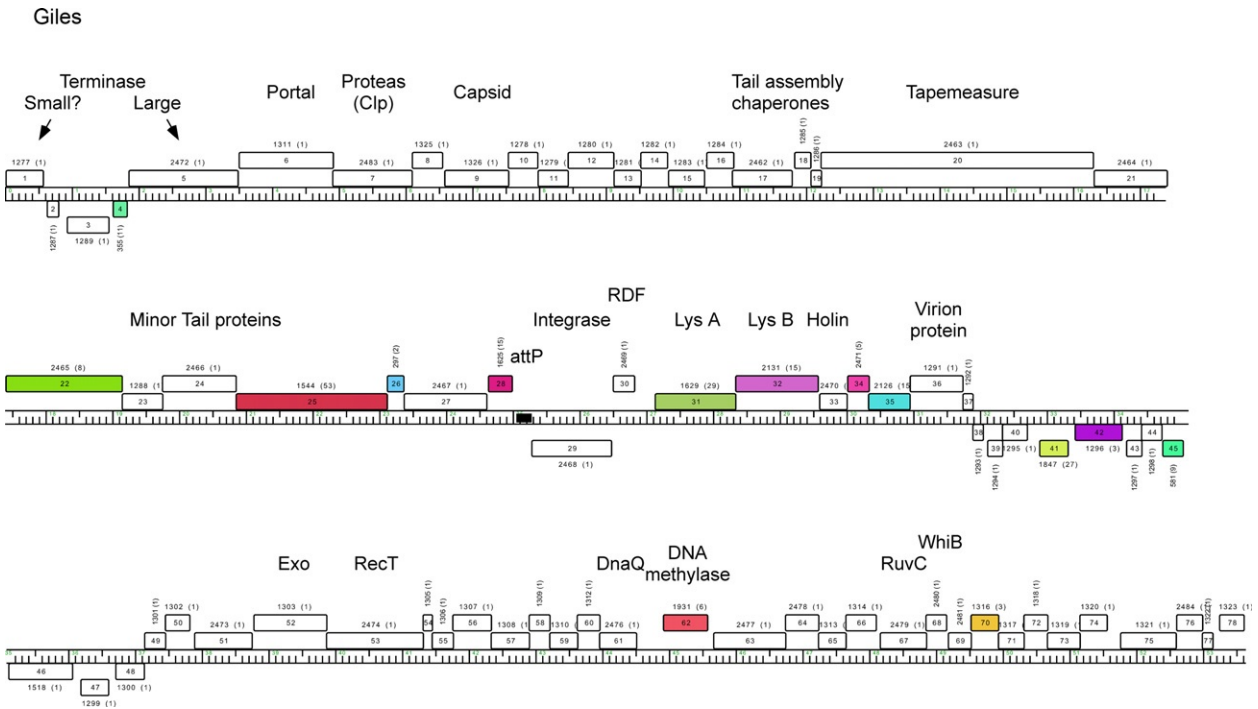


FIGURE 17 Map of the singleton phage Giles genome. See Figure 3A for further details on genome map presentation.

Author's personal copy

this position by an errant recombination event encoded by the integrase protein (Morris *et al.*, 2008). To the right of the virion structure and assembly operon there is one leftward-transcribed operon and one to the right. The leftward operon contains genes *38–48*, all of which are of unknown function. The rightward operon contains genes *49–78* and although the vast majority of these are orphams and have no known function, several do have potential functions of interest. For example, genes *52* and *53* encode putative exonuclease and RecT proteins, respectively, and presumably mediate homologous recombination events; this is supported experimentally (van Kessel and Hatfull, 2008a). Gene *61* encodes a DnaQ-like enzyme, a common function among many mycobacteriophage genomes, although the diversity of the encoded proteins is very high, and the closest homologue of Giles gp61 is a related gene encoded by *Kineococcus radiotolerans* (30% amino acid sequence identity). Gene *62* encodes a putative DNA methylase, although its function is unclear; DNA methylases can be components of restriction–modification systems, although if this were the case in Giles, it is unclear which gene might encode the putative restriction function. Gene *67* encodes a RuvC-like Holliday Junction resolvase, extending this as one of the most common functions encoded by mycobacteriophage genomes, albeit through the use of different classes of genes (i.e., RuvA, RusA, EndoVII); gene *68* encodes a WhiB-like gene regulator.

Although it was reported initially that the Giles genome was 54,512 bp in length (Morris *et al.*, 2008), this includes a segment at the extreme right end that was included due to an assembly error. Correction of the sequence generates a 53,746 bp genome, and the putative *metE*-like gene reported as Giles gene *79* is not actually part of the Giles genome (Hatfull *et al.*, 2010).

3. Wildcat

The singleton Wildcat forms plaques on *M. smegmatis* that are not evidently turbid, but not completely clear, and stable lysogens have not been reported. There is also no evidence for prophage stabilization functions in the genome, such as integrase or partitioning functions, nor are there any obvious candidates for a phage repressor. It does not infect *M. tuberculosis*. The genome is 78.3 kbp in length and contains defined cohesive termini with 11-base 3′ ssDNA extensions (Table I). A map of the Wildcat genome organization is shown in Figure 18. As with other singleton phages (and Clusters J and L for which only a single published genome is discussed here), there is a very high proportion (84%) of orphams.

Wildcat gene *26* encodes the putative terminase large subunit and is located over 8 kbp away from the physical left end of the genome. Immediately to its left are two other rightward-transcribed genes (*24* and *25*) of unknown function, although one of these could plausibly

Author's personal copy

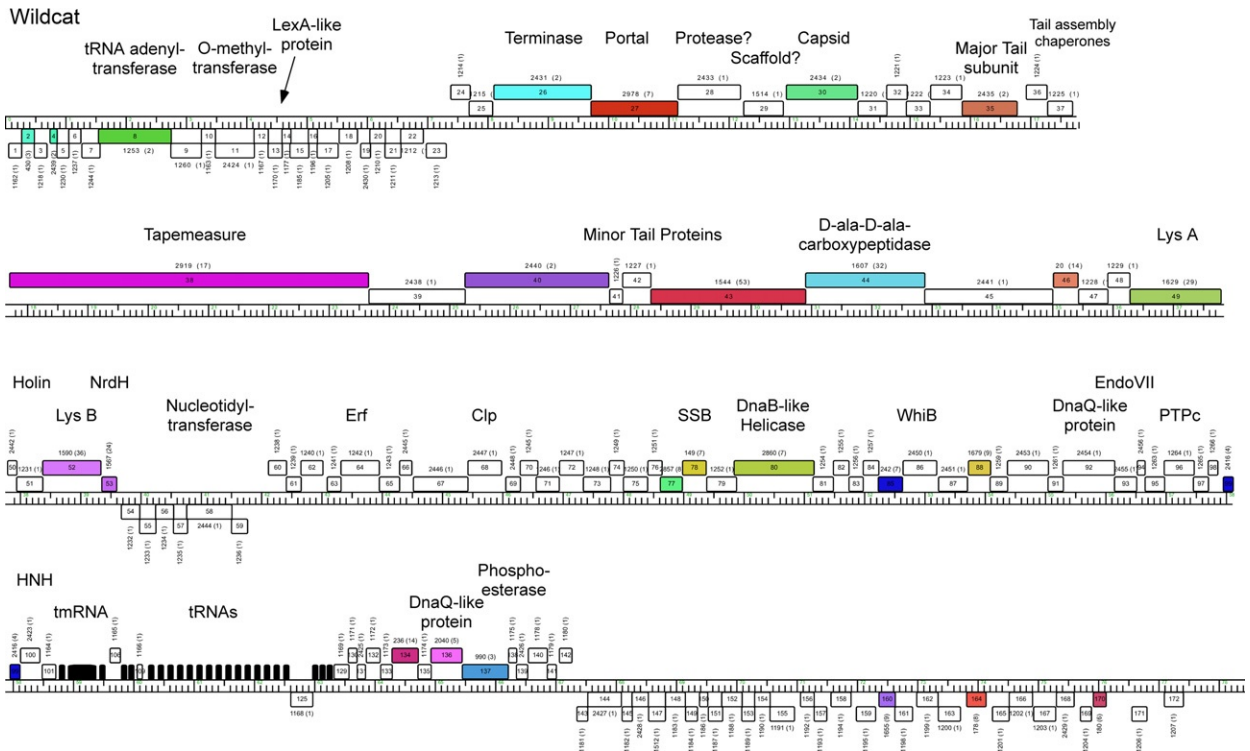


FIGURE 18 Map of the singleton phage Wildcat genome. See Figure 3A for further details on genome map presentation.

Author's personal copy

encode a terminase small subunit. Between the left end and gene *24* is a leftward-transcribed operon (genes *1–23*) containing mostly genes of unknown function. The presence of ''additional'' genes in this part of the genome (it is more typical for terminase genes to be close to their sites of action, i.e., near the genome end) extends a theme observed in the genomes of Clusters A, B, D, and Corndog. Three of these genes can be assigned putative functions. Gene *13* encodes a putative LexA-like transcriptional regulator, and gene *11* encodes a putative *O*-methyltransferase, a function seen in other genomes, including Omega and Corndog, although Wildcat gene *11* is quite different in sequence to these. Wildcat gene *8* encodes putative tRNA adenyltransferase activity, presumably involved in CCA addition to tRNAs (Fig. 18). The only other mycobacteriophage with a similar function is Myrna (gp28), although it shares no more similarity to Myrna gp28 (~35% amino acid identity) than it does to host PncA proteins. Both Wildcat and Myrna encode a large number of tRNAs, which may belie the requirement to encode this function, although we note that Cluster C1 phages also encode a large number of tRNAs but appear to lack such an activity.

The virion structure and assembly operon (genes *26–45*) is fairly canonical with uninterrupted synteny, and genes encoding putative terminase large subunit (*26*), portal (*27*), capsid (*30*), major tail subunit (*35*), tail assembly chaperone expressed via a programmed translational frameshift (*36* and *37*), tapemeasure (*38*), and minor tail proteins (*39–45*) can be predicted with confidence; gene *28* is a distant relative of LeBron gene *6* and likely encodes a protease, and gene *29*—a distant relative of LeBron gene *7*—is a strong candidate for encoding a scaffold protein in light of its location within the operon (Fig. 18). Wildcat gp44 has putative D-alanyl-D-alanine carboxypeptidase activity common to that of β-lactamase enzymes, which is commonly encoded by genes located among other tail genes, as in Subcluster A1, Clusters C, D, E, J, and singleton Corndog. The role of these putative proteins is unknown, but they presumably are involved in either cell wall binding or facilitating receptor recognition (see Section VI.A). The lysis cassette lies to the right of the virion structure and assembly operon and includes lysin A (*49*) and lysin B (*52*) genes and a putative holin gene (*51*) located between them. Immediately to the right of the lysis cassette is a small gene encoding a putative NrdH-like glutaredoxin. The role for such a redoxin is unknown, but genes with related functions are found in a number of other mycobacteriophages and there are many examples of them located in a similar position, just downstream of the lysis cassette. Examples are found in Cjw1, Omega, and LeBron, but in Cluster A and K phages it is encoded elsewhere in the genome. To the right of the lysis cassette are three apparent operons: two transcribed leftward (genes *54–59*; genes *143–172*) flanking a rightward operon (genes *60–142*). The two leftward-transcribed operons are virtually devoid

Author's personal copy

of any genes with predicted functions, the exception being gene *58* encoding a putative nucleotyltransferase. Only three of this entire repertoire of genes (*160*, *164*, and *170*) have homologues in other mycobacteriophages. The rightward operon contains several genes of interest, including the putative recombinase Erf (gene *64*), a Clp protease (gene *68*), SSB (gene *78*), a DnaB-like helicase (gene *80*), a WhiB-like regulator (gene *86*), two DnaQ-like but distantly related proteins (encoded by genes *92* and *136*), a putative PTPc-like phosphatase (gene *96*), and a putative phosphoesterase of unknown specificity (gene 137).

This operon also contains an impressive array of tRNA genes (23 in total), as well as a tmRNA gene. Although several proposals have been presented to explain the potential roles for mycobacteriophage-encoded tRNAs (Hassan *et al.*, 2009; Kunisawa, 2000; Sahu *et al.*, 2004), the variety of their numbers and types in mycobacteriophages is amazing (Table I). In addition to Wildcat, all Cluster C phages also have a large number of tRNA genes, Cluster E phages have two, Subcluster K1 phages have one, just one of the Cluster B phages has one (Nigel), and Subclusters A2 and A3 have between one and five. Thus for some phages it appears advantageous to have virtually a complete coding set of tRNA genes, whereas others appear to require no tRNA genes at all. Wildcat is the only phage outside of Cluster C that also encodes a tmRNA. It seems plausible that the phage-encoded tmRNA may serve to increase the efficiency of release of ribosomes from broken or otherwise damaged mRNAs, optimize translation efficiencies by maximizing the size of the pool of available ribosomes, or monitor protein folding (Hayes and Keiler, 2010). tRNA genes may also play a general role in enhancing the frequency of translation, although we cannot rule out that at least some of the tRNAs may be involved in the introduction of noncanonical amino acids into proteins.

In summary, Wildcat is certainly quite a wild phage with numerous features of interest that deserve a more detailed investigation. As with other singletons, annotation and interpretation of the genome lacks from the insights provided by the availability of more closely related phages.

IV. MYCOBACTERIOPHAGE EVOLUTION: HOW DID THEY GET TO BE THE WAY THEY ARE?

The collection of sequenced mycobacteriophage genomes —with its clusters of closely related phages, as well as those that are distantly related— provides abundant insights into their evolution, and viral evolution in general. The most prominent feature of their genomic architectures is that they are mosaic, that is, that the structure of each genome can be explained as being constructed from a set of modules that are being exchanged among the phage population (Hatfull, 2010; Hatfull *et al.*, 2006, 2008;

Author's personal copy

Pedulla *et al.*, 2003). These modules—composed of either single genes or larger groups of genes—are thus located in different phage genomes that are otherwise not closely related. This mosaicism can be seen at two different levels of comparative genomic analyses. When genomes are compared at the nucleotide sequence level, relatively recent exchange events can be seen, and several examples have been described previously (Hatfull *et al.*, 2010; Pope *et al.*, 2011). However, the pervasive nature of the genomic mosaicism is manifested by looking at protein sequence comparisons because shared gene ancestries can be detected even when they have diverged sufficiently long ago in evolutionary time that nucleotide sequence commonality is no longer evident (Pedulla *et al.*, 2003). Representing genome mosaicism through a comparison of standard approaches such as phylogenetic trees is made complicated by the fact that the component genes may be resident in entirely different genomes (Hatfull *et al.*, 2006). An alternative method of representation uses phamily circles in which each of the genomes in the analysis is placed around the circumference of a circle and an arc is drawn between those genomes that share a gene member of a particular phamily of related genes (Hatfull *et al.*, 2006). Examples in Figure 19 show phamily circles for five of their eight consecutive genes (*21*–*28*) in the Che9c genome; the remaining three genes are orphams and thus no related mycobacteriophage genes have yet been identified. One of these genes (Che9c *22*) has only a single related gene (forming Pham274), which is located in the Subcluster A3 genome, Bxz2 (Fig. 19). Che9c gene *24* is related to 22 other genes in Pham1628 and is found in a variety of genomes in Clusters A, F, I, J, and K, but not in Bxz2 (Fig. 19). Che9c genes *25*, *26*, and *27* (members of Phams 1584, Pham1758, and Pham 1451, respectively) have relationships that are different yet again. Thus all the individual genes in this region appear to have distinct evolutionary histories and arrived at their genomic locations through different evolutionary journeys (Fig. 19). This mosaicism complicates greatly the task of constructing whole genome phylogenies, as the genome relationships are fundamentally reticulate in nature (Lawrence *et al.*, 2002; Lima-Mendez *et al.*, 2007).

A hallmark feature of phage genome mosaicism is that in cases where recombination events can be inferred, they occurred at gene boundaries, or occasionally at domain boundaries. Usually this is observed where closely related phage genomes have undergone relatively recent recombination events and genome discontinuities can be seen at the nucleotide level (Hatfull, 2008; Pope *et al.*, 2011). In such cases, it is not uncommon for sequence discontinuity to appear precisely at or very close to the start and/or stop codons of genes (Hendrix, 2002). This could occur either by a process of targeted recombination events or as a consequence of functional selection from a large number of possible exchange events, most of which generate nonviable progeny and were subsequently lost from the

Author's personal copy

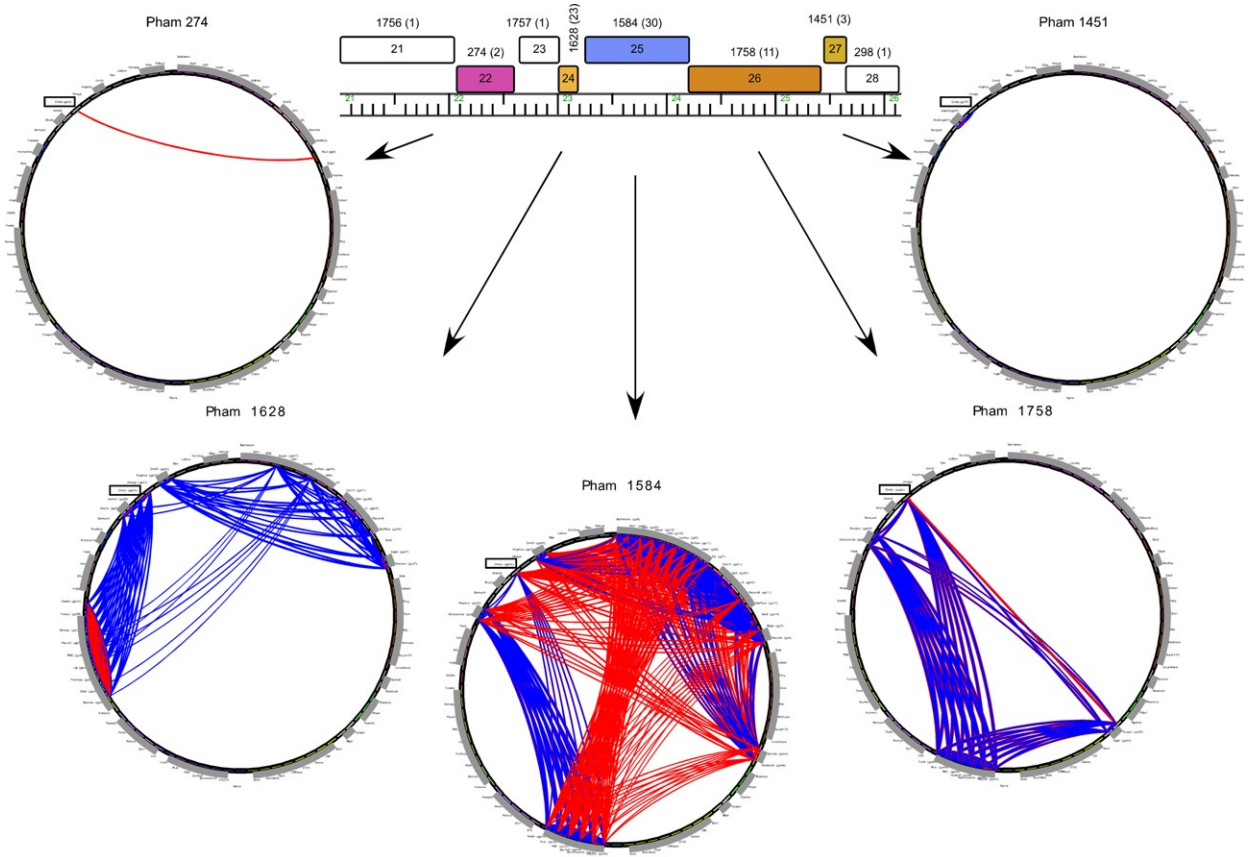


FIGURE 19 Genome mosaicism in the Che9c genome. A segment of the Subcluster I2 Che9c genome encoding genes *21–28* is shown, as described for Fig. 12. Genes *21*, *23*, and *28* are single members of orphams and thus are shown as white boxes. Genes *22* and *24–27* each have relatives in other mycobacteriophage genomes; these are represented as phamily circles for the five respective phams. In each phamily circle, all

Author's personal copy

population (Hendrix *et al.*, 1999). It should be noted that interpreting where recombination events occur is complicated by the fact that subsequent rearrangement events could have occurred between the genomes being compared (Pedulla *et al.*, 2003). In addition, it is impossible to know what specific recombination events gave rise to the numerous mosaic relationships revealed only through amino acid sequence comparisons.

How does genome mosaicism arise? First it is helpful to note that while mutational changes involving nucleotide substitutions clearly occur and are an important component of phage evolution, this does not contribute directly to genome mosaicism, and acquisition of genome segments from other contexts—either phage or host—by horizontal genetic exchange offers a more general explanation (Hendrix *et al.*, 1999, 2000). Homologous recombination between genome segments with extensive sequence similarity also plays an important role in genome evolution in that it can generate new combinations of gene content, but does not—with the exception of the process described later—create new gene boundaries that are the key to juxtaposing one module next to another.

Four known mechanisms are likely to make substantial contributions to the creation of new module boundaries, although their relative importance is ill-defined. The first is the process of homologous recombination events occurring at short conserved sequences at gene boundaries (Susskind and Botstein, 1978). This process has been proposed in other phages (Clark *et al.*, 2001), and there are a few examples in which this could have played a role in mycobacteriophage mosaicism (Pope *et al.*, 2011). We also note that the 13-bp stoperator sites present in Cluster A genomes, which are predominantly located near gene boundaries, could be ideal targets for such targeted recombination events (see Sections III.A and V.A.1). For the most part, however, short conserved boundary sequences are not obvious at most of the mosaic boundaries that can be identified (Pedulla *et al.*, 2003). However, most of these are revealed through amino acid sequence comparisons and occurred long ago in evolutionary time such that any conservation at the boundaries would

---

80 genomes are represented (in the same order and grouped according to cluster/subcluster) around the circumference of the circle, and an arc is drawn between those members of the genomes containing a gene that is a member of that pham. Red and blue arcs show BlastP and ClustalW comparisons, respectively, and the thickness of the arc reflects strengths of the relationships. The position of Che9c is boxed in each circle. In Pham 1451 there are only two relatives present in the Subcluster I1 genomes. In Pham 274, there is only a single relative that is in the unrelated Subcluster A3 genome, Bxz2. Phams 1628, 1584, and 1758 each have multiple members but distributed among different clusters and subclusters. This suggests that each of the eight Che9c genes, *21–28*, have arrived in Che9c through distinct evolutionary journeys. This mosaicism is a hallmark of bacteriophage genomic architectures.

Author's personal copy

have been long lost. The second process is site-specific recombination events in which secondary sites have been used by a site-specific recombinase to give rise to insertions in atypical locations. Although this is unlikely to be a predominant process, phages often encode site-specific recombinases, including both tyrosine- and serine-family integrases. One notable example is observed in phage Giles (see Section III.M.2), in which the integration cassette is located among the tail genes, and could have moved there through integrase acting at a secondary site within the virion structure and assembly operon (Morris *et al.*, 2008). The third process is by movement of mobile elements such as transposons and other mobile elements such as inteins, homing endonucleases, and introns, generating both insertions into new genomic locations, and by transposase-mediated rearrangements such as adjacent deletions and inversions. Transposons are not common in mycobacteriophages but several have been recognized. The strongest evidence is for MPME elements found in Cluster G and many Cluster F genomes (see Sections III.G and III.F); these are clearly involved in interrupting what are otherwise conserved gene syntenies (Sampson *et al.*, 2009). Another example is the putative IS110 family insertion sequence present in Omega and its distant relative in Bethlehem (see Sections III.A and III.J). Numerous examples of inteins and HNH-like homing endonucleases exist throughout genomes, but no mycobacteriophage introns have been described. The fourth—and probably the most important contributor—is illegitimate or nontargeted recombination processes that occur without requirement for extensive sequence identity (Hendrix, 2003; Hendrix *et al.*, 1999; Pedulla *et al.*, 2003). It is unclear what mediates such events, although it is noteworthy that bacteriophages commonly encode their own general recombinases, such as phage λ Red systems, RecET-like systems, and P22 Erf-like systems. Mycobacteriophages are no exception, and there are now many examples of RecT-like recombinases (associated with several different types of exonuclease, some of which are related to RecE and some which are not), Erf-like functions, and RecA-like proteins. Examples of some of these phage-encoded recombinases are known to mediate recombination over shorter segments of sequence identity than is typically favored by host recombination systems; they can also tolerate substantial differences between recombining partners (Martinsohn *et al.*, 2008). Although the efficiency of recombination at ultrashort sequence commonalities (such as codons or ribosome-binding sites) is expected to occur at very low frequencies, and multiple events may be required to generate viable progeny, with a potentially long evolutionary history (2 to 3 billion years?) and a high incidence of infection (estimated to be about 10^{23} infections per second globally), inefficiency is unlikely to be an impediment to generating the extent of mosaicism seen in the phage population today. It should also be noted that such illegitimate recombination events are likely to occur more

Author's personal copy

frequently between phages and their host genomes that are often 100 times larger, consistent with the common finding of host genes in phage genomes (Pedulla *et al.*, 2003). There are numerous examples of genes present within mycobacteriophages that are not typically present in phage genomes, with the queuosine biosynthesis genes in Rosebush (see Figs. 4 and 5) being a good example (Pedulla *et al.*, 2003). Finally, we note that generating new gene boundaries either by transposition or by illegitimate recombination is a highly creative process in that DNA sequence elements can be placed together in combinations that did not exist previously in nature. Although most illegitimate recombination events are expected to make genomic trash, the process is one of very few that can create entirely new types of genes.

Comparative genomic analysis of SPO1-like phages led to the suggestions that newly acquired genes are, on average, relatively small (Stewart *et al.*, 2009), and a similar conclusion arises from a comparison of mycobacteriophage genomes (Hatfull *et al.*, 2010). This is consistent with the predominant role of illegitimate recombination because most events are likely to occur within reading frames, and thus selection for function is expected to drive toward functional domains rather than multidomain proteins. This could also account for the reason that phage genes are, on average, only about two-thirds the average size of host genes (Hatfull *et al.*, 2010).

V. ESTABLISHMENT AND MAINTENANCE OF LYSOGENY

Temperate phages are of particular interest for a variety of reasons. For example, they typically employ gene regulatory circuits that can provide insights into novel systems for gene expression and control, as well as being potentially useful for genetic manipulation of the host. Similarly, phage integration provide insights into mechanisms of site-specific recombination and how directionality is controlled, as well as providing the basis for novel plasmid vectors for host genetics (Hatfull, 2010). Temperate phages also often carry genes expressed from the prophage state and contribute to lysogenic conversion of the physiological state of the host. All of these aspects are applicable to mycobacteriophages, and the intimacy of phage–host relationships inherent in temperate phages is particularly intriguing.

A. Repressors and immunity functions

1. Cluster A immunity systems

Genes encoding phage repressors have been identified in remarkably few mycobacteriophages, and there is no complete understanding of life cycle regulation in any of them. Perhaps the best studied are the immunity

Author's personal copy

systems of mycobacteriophage L5—and its unsequenced but closely related phage L1 (Subcluster A2)(Lee *et al.*, 1991)—where the repressor has been identified and characterized (Bandhu *et al.*, 2009, 2010; Brown *et al.*, 1997; Donnelly-Wu *et al.*, 1993; Ganguly *et al.*, 2004, 2006, 2007; Nesbit *et al.*, 1995; Sau *et al.*, 2004)(see Section III.A). A number of other phages encode related repressors, including other Cluster A members, and the Cluster C phage, LRRHood, and the Cluster F phage, Fruitloop, although they are diverse at the sequence level, and pairwise relationships between repressors from different Subclusters in Custer A can be below 30% amino acid sequence identity. The Subcluster A1 phage Bxb1 repressor is the only other one that has been analyzed in any detail (Jain and Hatfull, 2000).

The L5 repressor (gp71) is a 183 residue protein containing a strongly predicted helix-turn-helix DNA-binding motif and was identified through two key observations. First, when the repressor gene is expressed in the absence of any other phage-encoded functions it confers immunity to superinfection by L5. Second, mutations in the repressor confer a clear plaque phenotype; point mutations in the repressor gene can lead to a temperature-sensitive clear-plaque phenotype and lysogens that are thermoinducible (Donnelly-Wu *et al.*, 1993). Unlike most other well-studied repressors, it is predominantly a monomer in solution and recognizes an asymmetric sequence in DNA (Bandhu *et al.*, 2010; Brown *et al.*, 1997). A primary target of regulation is the early lytic promoter P_{left}, which is situated at the right end of the genome and transcribed leftward (Fig. 3A). The L5 P_{left} promoter is highly active and contains −10 and −35 sequences corresponding closely to the consensus sequences for *E. coli* sigma-70 promoters (Nesbit *et al.*, 1995). There are two 13-bp repressor-binding sites at P_{left}, one of which (site 1) overlaps the −35 sequence and the other (site 2) is located ∼100 bp downstream within the transcribed region. L5 gp71 binds to these two sites independently, and binding to site 2 does not substantially influence repression of P_{left}; when gene *71* is provided on an extrachromosomal plasmid, P_{left} is downregulated about 50-fold (Brown *et al.*, 1997) through repressor binding to site 1. The binding affinity of gp71 for site 1 is modest, with a K_d of about 5×10^{-8} *M*, and binding to site 2 is about 5- to 10-fold weaker (Brown *et al.*, 1997). Interestingly, repression by binding at site 1 may not be mediated by promoter occlusion, but rather by RNA polymerase retention at the promoter. This is indicated by the observation that phage mutants can be isolated [designated as class III (Donnelly-Wu *et al.*, 1993)] that have mutations within the repressor gene but have a dominant-negative phenotype, being competent to infect a repressor-expressing strain. Such gp71 variants could thus bind to site I without retaining RNA polymerase and prevent that action of wild-type gp71. A surprising observation was that the L5 genome contains a large number of potential repressor-binding

Author's personal copy

sites located throughout the genome (see Section III.A). Initially, a total of 30 putative sites were identified, 24 of which (including sites 1 and 2) were shown biochemically to be bound by gp71 (Brown *et al.*, 1997). These sites conform to the asymmetric consensus sequence 5′-GGTGGMTGTCAAG (M is either A or C), where eight of the positions are absolutely conserved and three others contain only a single departure from the consensus (Brown *et al.*, 1997); roles for the six nonbinding sites cannot be ruled out, as weaker gp71 association may be biologically relevant. Sites are not positioned randomly in the genome but have two important features in common. First, they are oriented in predominantly just one direction relative to the direction of transcription. Thus of the five sites located within the left arm (between the physical left end and the integrase gene; Fig. 3), four (sites 20–23) within the predicted rightward-transcribed region (genes *1–32*) are oriented in the ''−'' direction; the other (site 24, located between the physical left ends and gene 1) is in the ''+'' orientation (Brown *et al.*, 1997). It is not known if this segment is transcribed or not. In the right arm (between the integrase gene and the physical right end, Fig. 3A), all of the sites in the leftward-transcribed region (genes *23–88*) are oriented in the ''+'' orientation. Second, sites are typically located within short intergenic intervals, often overlapping the putative start and stop codons of adjacent genes. When one or more of these binding sites is inserted between a heterologous promoter (hsp60) and a reporter gene (*FFLux*), binding of gp71 has a polar effect on gene expression. This effect is repressor dependent, is strongly influenced by orientation of the site relative to transcription, and is amplified by the presence of multiple sites (Brown *et al.*, 1997). Because repressor binding appears to prevent transcription elongation rather than initiation, these sites (other than site 1) are referred to as ''stoperators'' (Brown *et al.*, 1997). The mechanism by which this occurs is unknown, but an attractive model is that the repressor interacts directly with RNA polymerase and perhaps retains it at the stoperator site, consistent with the model for action as a repressor by RNA polymerase retention at site 1. It is postulated that these sites play a role in silencing the L5 prophage, ensuring that phage genes potentially deleterious to growth of a lysogen are not expressed from errant transcription events during lysogeny (Brown *et al.*, 1997). However, there is not yet any formal demonstration that additional promoters are not overlapping all or some of these sites or that removal of any of these sites influences either prophage stability or fitness of L5 lysogens.

The regulation of L5 gene *71* is poorly understood, although there is a set of three putative promoters located upstream in the gene *71–72* intergenic region that are presumably responsible for gp71 synthesis from a prophage. The reason for three promoters is unclear. Curiously, even though these three promoters are downregulated during lytic growth, evidence shows that the repressor gene is transcribed during early lytic

Author's personal copy

growth from transcripts arising from P_{left} (Fig. 3)(Nesbit *et al.*, 1995). Presumably, other phage-encoded functions prevent gp71 from acting during lytic growth, although none have been identified. Another conundrum arises because the three promoters upstream of gene *71* are active in a nonlysogen, such that although it is simple to model how lysogeny is established, it is less easy to imagine how lytic growth ensues after infection. Presumably, either the action of gp71 itself is modulated — perhaps either by post-translational modification or by degradation—or a second regulator prevents expression during the establishment of lytic growth. Although no additional L5 genes have been specifically identified as playing a role in the L5 lytic–lysogenic decision, clear plaque mutants have been identified with reduced frequencies of lysogeny (similar to cIII mutants of phage λ), and genes located within the region to the right of gene *71* are implicated (Donnelly-Wu *et al.*, 1993; Sarkis *et al.*, 1995). Finally, we note that L5 lysogens are not strongly inducible by DNA-damaging agents, even though this is a common feature of many other temperate phages. Lysogens do undergo spontaneous induction to release particles into the supernatant of a liquid culture, but the nature of repressor loss-of-function is not known.

Mycobacteriophage Bxb1 encodes a related repressor (gp69), although it shares only 41% amino acid identity with L5 gp71 and the two phages are heteroimmune (Jain and Hatfull, 2000; Mediavilla *et al.*, 2000). However, there are many common features of the two immunity systems, including multiple promoters upstream of the repressor gene [two in Bxb1 (Jain and Hatfull, 2000)], a repressor-regulated early lytic promoter, and multiple stoperator sites located throughout the genomes. In Bxb1 there are 34 putative 13-bp asymmetric stoperator sites, corresponding to the consensus 5′-GTTACGWDTCAAG (W is A or T), with notable differences from the L5 consensus at positions 1, 4, and 5. Most of these share the same features of the L5 stoperators in being located within short intergenic regions and oriented in one direction relative to the direction of transcription. Bxb1 gp69 binds with a similar affinity to its binding site as L5 gp71 does to its sites, but recognition of each other's sites occurs only at a much lower affinity (~1000-fold lower), accounting for their heteroimmune phenotype.

Prior to its genomic characterization, mycobacteriophage D29 was thought to be a substantially different phage than L5 and others, partly because it forms completely clear plaques and partly because it infects *M. tuberculosis* readily [L5 also infects *M. tuberculosis*, but has specific requirements for high calcium concentrations that D29 does not (Fullner and Hatfull, 1997)]. Genomic analysis showed that it is a derivative of a temperate parent that has suffered a 3.1-kbp deletion removing the repressor and several closely linked genes (Ford *et al.*, 1998a). The deletion event likely occurred relatively recently—perhaps at the time of its

Author's personal copy

isolation (see Section III.A)—and D29 is subject to gp71-mediated L5 immunity (Ford *et al.*, 1998a). Most of the stoperator sites identified in L5 are present at similar positions in L5, and the 13-bp consensus sequence is the same as that of L5.

More recently, the bioinformatic analysis of immunity specificities has been extended to all of the known Cluster A phages and concludes that these specificities closely mirror the subcluster divisions. That is, all of the phages within a subcluster form a homoimmune group, but none offers immunity to phages from other subclusters (Pope *et al.*, 2011). Although several other Cluster A phages appear to contain defective repressors such that stable lysogens cannot be recovered, all contain predicted stoperator sites varying in number from 23 in Che12 and Bxz2 to 36 in Jasper (Pope *et al.*, 2011). From all 17 Cluster A phages, a total of 453 potential sites have been identified, and although only those in L5 and Bxb1 have been shown to be true binding sites, some general features are evident. In particular, positions 1 and 13 are absolutely conserved (G in both positions) and positions 9 and 12 are highly conserved (T and A, with nine and two departures, respectively). Positions 2 through 6 maybe the primary determinants of specificity and can likely be discriminated by differences in the second helix of the HTH motifs of the repressors (Pope *et al.*, 2011), although this awaits detailed experimental analysis. DNA protection and mutational analysis of L1 repressor binding are consistent with the bioinformatic findings (Bandhu *et al.*, 2010).

Mycobacteriophage Fruitloop and LRRHood—members of Clusters F and C, respectively—both contain genes related to the Cluster A repressors, even though stable lysogens have not been reported for either phage. Comparisons of repressor genes show that these are very closely related to the repressor of Cluster A1 phages (Fig. 20), and LRRHood gp44 has only a single amino acid departure from Bxb1 gp69 (Pope *et al.*, 2011). However, neither LRRHood nor Fruitloop contains multiple binding sites related to the Bxb1 stoperators. There is not a single potential repressor-binding site in the LRRHood genome, and in Fruitloop there is just a single site located upstream of gene *39* that could play a role in autoregulation. As discussed previously, the presence of repressor genes in these phages could have been selected to confer protection to either lysogens or infected cells from superinfection with Bxb1-like Cluster A1 phages (see Section III.F).

2. Cluster G immunity systems

Putative repressor genes have also been identified in Cluster G phages, such as BPs and its closely related relatives Halo, Angel, and Hope (Sampson *et al.*, 2009). These phages are temperate, and stable lysogens can be recovered from infected cells. The BP repressor (gp33) is not closely related to Cluster A-encoded or any other phage repressors but does

Author's personal copy

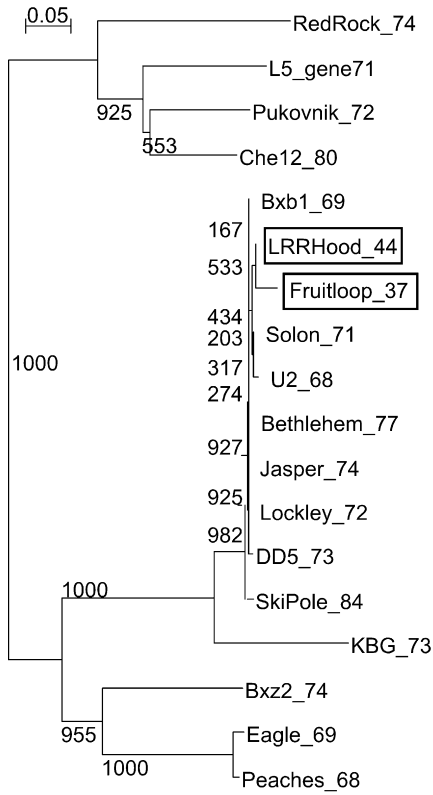


FIGURE 20 Phylogenetic tree of mycobacteriophage repressor proteins. The neighbor-joining phylogenetic tree of mycobacteriophage repressor-like proteins was generated from an alignment created in Cluster X and drawn using NJPlot. All of the repressors shown are encoded by Cluster A phages, with the exception of LRRHood and Fruitloop (boxed), which are members of Subclusters C1 and F1, respectively. THe LRRHood and Fruitloop repressors are related more closely to those in Cluster A1 (i.e., Bxb1 and its relatives) than to others.

contain a putative helix-turn-helix DNA-binding motif, and expression of gp33 confers immunity to superinfection by all of the Cluster G phages (see Section III.G). The repressor gene is located immediately upstream of the integrase gene (*32*) and the two genes are predicted to overlap. A notably unusual feature of genome organization is that the crossover site for integrative recombination at *attP* is located within the repressor gene itself, such that the gene product expressed from the prophage is 33 residues shorter than the virally encoded form. This suggests the possibility that integration plays a central regulatory role in the lytic–lysogenic decision.

Author's personal copy

B. Integration systems

Phages within Clusters A, E, F, G, I, and K and singletons Giles, Omega, and LeBron all encode integrases, mostly of the tyrosine-recombinase family. However, several phages—all within the Cluster A—encode serine-integrases, including all of Subcluster A1, Subcluster A3, and Peaches of Subcluster A4. Although most of the Subcluster A2 phages encode tyrosine-integrases, an interesting exception is RedRock, which encodes putative ParA and ParB proteins at the same genomic location as its Subcluster A2 relative encode tyrosine-integrases (Fig. 3B).

1. Tyrosine-integrase systems

The best studied of mycobacteriophage tyrosine-integrases is that encoded by L5. L5 integrase is a distant relative of the phage λ prototype, but shares many central features. For example, L5 gpInt (gp33) contains two DNA-binding specificities: one encoded in a small (65 residue) N-terminal domain that binds to arm-type sites in *attP* and a second within the larger C-terminal domain that recognizes core-type sequences in *attP* and *attB* (Peña *et al.*, 1997). Amino acid residues critical for the chemistry of strand exchange, including catalytic tyrosine, are all well conserved. In the L5 genome, the *attP* site is located to the 5′ side of the integrase gene and is ~250 bp long, containing core-type integrase-binding sites flanked by arm-type integrase-binding sites (Peña *et al.*, 1997); the *attB* site overlaps a $tRNA^{gly}$ gene in the *M. smegmatis* genome (Lee and Hatfull, 1993; Lee *et al.*, 1991). L5 integrase-mediated integrative recombination requires L5 gpInt, a host-encoded mycobacterial integration host factor (mIHF), and *attP* and *attB* DNAs (Lee and Hatfull, 1993; Pedulla *et al.*, 1996). DNA supercoiling stimulates recombination *in vitro*, but this is observed if either of the DNA molecules is supercoiled (Peña *et al.*, 1998). The host factor mIHF is quite distinct from other IHF-like proteins, and its name reflects its function rather than any sequence or structural similarity (Pedulla *et al.*, 1996). It contains a single subunit with DNA -binding properties, is an essential gene in *M. smegmatis* (Pedulla and Hatfull, 1998), but does not appear to bind either *attP* or *attB* with any specificity. Nonetheless, it strongly promotes the formation of stable tertiary complexes containing gpInt, mIHF, and *attP* DNA (Pedulla *et al.*, 1996). Interestingly, there appear to be alternative pathways for the assembly of synaptic complexes (Peña *et al.*, 2000) containing *attB*, and within which strand exchange occurs. Cleavage occurs seven bases apart within the core region to generate 5′ extensions, and cleavage is associated with covalent linkage of gpInt to the 3′ ends of the DNA (Peña *et al.*, 1996); the 7-bp overlap region corresponds to the anticodon loop of the $tRNA^{gly}$ gene at *attB*. The directionality of L5 integrase-mediated recombination is determined by recombination directionality factor

Author's personal copy

gp36 (gpXis)(Lewis and Hatfull, 2000). L5 gp36 is small (56 residues) and binds to four putative-binding sites in *attP* and *attR* (Lewis and Hatfull, 2003). L5 gp36 is proposed to impart a substantial DNA bend at these sites and thus dictates the ability of integrase to form recombinagenic protein–DNA complexes (Lewis and Hatfull, 2003). When bound to *attR* it promotes formation of a complex in which gpInt is bound simultaneously to the core and arm-type sites in *attR* to form an intasome that can synapse with at *attL*–intasome (Lewis and Hatfull, 2003). L5 gp36 thus strongly stimulates excisive recombination. In contrast, when L5 gp36 is bound to *attP* DNA, it discourages formation of an intasome-like structure that can synapse with *attB* DNA, which inhibits integrative recombination (Lewis and Hatfull, 2003).

2. Serine-integrase systems

Serine-integrases are unrelated to tyrosine-integrases and typically contain an N-terminal domain of 140–150 residues related to the catalytic domain of transposon resolvases such as Tn*3* and γδ and a large C-terminal domain with DNA-binding activity (Smith and Thorpe, 2002). The best-studied of the mycobacteriophage-encoded systems is that of Bxb1, although the related system encoded by the prophage-like element, ϕRv1, has also been investigated. Bxb1 gpInt (gp35) is a 500 residue two-domain protein that catalyzes site-specific recombination between an *attP* site located to the 5′ side of the gene *35* and an *attB* site located with the *M. smegmatis groEL1* gene (Kim *et al.*, 2003). Both *attP* and *attB* are small, and the minimally required sites contain 48 and 38 bp, respectively (Ghosh *et al.*, 2003). Strand exchange occurs at the centers of these sites, and Bxb1 gpInt cleaves to generate two-base 3′ extensions; strand exchange involves the formation of gpInt–DNA covalent linkages with the serine at position 10 linked to the 5′ ends of the DNA (Ghosh *et al.*, 2003). Bxb1 gp35 efficiently mediates site-specific recombination between *attP* and *attB in vitro* to generate *attL* and *attR*, and no additional proteins are required (Ghosh *et al.*, 2003; Kim *et al.*, 2003). The reaction is not stimulated significantly by DNA supercoiling and does not require the addition of either metal ions or high-energy cofactors (Ghosh *et al.*, 2003). This reaction is strongly directional, and Bxb1 gpInt alone does not catalyze recombination between *attL* and *attR*; it also fails to catalyze recombination between any pair of sites other than between *attP* and *attB*. The simplicity of this reaction greatly facilitates biochemical dissection of the reaction, with the origins of the site specificity and the control of directionality as central questions of interest.

Both *attP* and *attB* are quasi-symmetric in nature, being composed of imperfectly inverted repeats flanking the 5′-GT central dinucleotide, and gpInt binds to each site as a dimer (Ghosh *et al.*, 2005). However, the P and P′ half sites in *attP* are distinctly different from the B and B′ *attB* half-sites, although all four half-sites contain a 5′ ACNAC motif in symmetrically

Author's personal copy

related positions (Ghosh *et al.*, 2003). These structures raise several interesting questions. First is the issue as to whether gpInt contains a single DNA recognition motif that somehow adapts to interact with the two different types of half-sites or whether there are two separate structural motifs, each capable of recognizing either B-type or B-type half-sites. Thus far there is no evidence for more than one type of DNA recognition motif, and the only mutants that discriminate between binding to *attP* and *attB* have substitutions in the putative linker region that joins the two domains (Ghosh *et al.*, 2005). A second issue is in regard to the relative orientation of synapsis, as each site is quasi-symmetrical, and presumably synapsis is mediated by protein–protein interactions between gpInt dimers bound to *attP* and *attB*. Interestingly, synapsis does indeed appear to occur in an orientation-independent manner, and it is only the asymmetric 5′-GT central dinucleotide that determines the orientation of integration (Ghosh *et al.*, 2003). Thus wild-type *attP* and *attB* sites can synapse in both possible orientations (these are referred to as parallel and antiparallel alignments, although the actual configurations are not known) with equal probabilities. In the productive orientation, one helix can rotate 180° around the other to generate a recombinant configuration within which religation to the partner DNA can proceed. In the nonproductive configuration, after 180° rotation of the helices, bases at the central dinucleotide are noncomplementary and ligation does not occur (Ghosh *et al.*, 2003). However, rotation can proceed for one or more subsequent rounds to realign the central nucleotide bases such that they are in the parental—and thus ligatable—position. Changing a single base in the central dinucleotide of both *attP* and *attB* such that the central nucleotides are palindromic thus leads to complete loss of orientation specificity, with approximately equal efficiencies of ligation of the P half-site with B and B′; likewise for P′ (Ghosh *et al.*, 2003). Site specificity for integrative recombination likely results from the specificity for synapsis. That is, even though integrase binds as a dimer to all four possible sites, *attP*, *attB*, *attL*, and *attR*, synapsis only occurs between gpInt-bound *attB* and *attP* complexes. The molecular basis for this is not known, but presumably gpInt adopts different conformations when bound to the four different sites, such that noncognate combinations are excluded conformationally. This raises the question as to how excision occurs, where gpInt bound to *attL* and *attR* must somehow presumably adopt conformations that are productive (Ghosh *et al.*, 2006). Genetic analysis identified a second phage-encoded protein, gp47, acting as an RDF in that it is required to enable integrase-mediated excisive recombination between *attL* and *attR*. Bxb1 gp47 also inhibits integrative recombination (Ghosh *et al.*, 2006), a common property of RDF proteins (Lewis and Hatfull, 2001). The molecular mechanism by which Bxb1 gp47 switches site specificity for gpInt is not known; it does not bind DNA, but rather associates with gpInt–DNA complexes and seems to do so differently depending on which type of site is bound

Author's personal copy

(Ghosh *et al*., 2006). This is at least consistent with a model in which gp47 modulates the conformation of gpInt, enabling productive configurations when it is bound to *attL* and *attR*, but not when it is bound to *attP* and *attB*. A notable consequence of the finding that *attP* and *attB* are essentially symmetrical for the purposes of synapsis is that *attL* and *attR*, both of which contain one B-type and one P-type half-site, are essentially identical (Ghosh *et al*., 2006, 2008). This predicts that asymmetry of the central dinucleotide again plays a critical role in determining productive recombination for excision, as gpInt bound to *attL* is expected to promote synapsis just as efficiently with itself as with gpInt bound to *attR*. This is confirmed experimentally, because switching the central dinucleotide of *attL* to make it palindromic is sufficient to generate an efficient three-component system requiring just gpInt, gp47, and the mutant *attL* site (Ghosh *et al*., 2008). It is also noteworthy that the asymmetry of *attL* and *attR* (each containing one B-type and one P-type half-site) is also reflected in the orientation of synapsis. Thus in each of the synaptic interactions observed, an gpInt protomer bound to a B-type half-site must interact with one bound to a P-type half-site (Ghosh *et al*., 2008).

Bxb1 255 residue gp47 is an unusual RDF and has no sequence similarity to other RDF proteins. It is not closely linked to the integrase gene as is often observed for RDF's, but is located approximately 5 kbp to its right, among genes predicted to be involved in DNA replication, including DNA polymerase and DNA primase genes (Ghosh *et al*., 2006). Strangely, there are relatives of Bxb1 gp47 in all 17 of the Cluster A phages, including all those that encode tyrosine-integrases, and indeed in L5 where all the phage-encoded genes required for efficient site-specific recombination—both integrative and excisive—are known (see Sections III.A and V.B.1 and Fig. 3B). The simplest interpretation is that Bxb1 gp47 is a dual function protein, fulfilling a common role of Cluster A phages—most likely in DNA replication—but also co-opted for use as an RDF in Bxb1. This raises the question as to whether homologues of Bxb1 gp47 also perform the RDF function in those phages that encode more distantly related serine-integrases, such as Bxz2 and Peaches (all of the serine-integrase encoded by Subcluster A1 phages are very similar to each other), or whether alternative proteins have been adopted (see Section III.A).

The serine-integrase system encoded by the *M. tuberculosis* prophage-like element ϕRv1 sheds some light on at least some of the questions raised by the Bxb1 system (Bibb *et al*., 2005; Bibb and Hatfull, 2002). Like Bxb1, requirements for integration *in vitro* are simple, requiring *attP* and *attB* partner DNAs, and ϕRv1 gpInt (Bibb and Hatfull, 2002). However, the reaction is somewhat slow and inefficient relative to the Bxb1 reaction. Interestingly, the ϕRv1 element integrates into a repetitive element in *M. tuberculosis* and is therefore found in several different chromosomal locations. The putative *attB* sites differ for each of the repeated sequences, although four of them are active as sites for recombination (Bibb and

Author's personal copy

Hatfull, 2002). The RDF protein for the ϕRv1 system has been identified, and the 73 residue protein (Rv1584c) is completely unrelated to Bxb1 gp47 (Bibb and Hatfull, 2002). Yet more surprising, the ϕRv1 RDF is related to Xis-like proteins associated with tyrosine-integrases, including L5 gp36 (Bibb and Hatfull, 2002). While the ϕRv1 RDF may have DNA-binding activity, this does not appear to be required for excision, as only the same minimal sequences are required for excision as they are for integration, all of which are apparently involved in close interactions with gpInt (Bibb *et al.*, 2005). It is thus likely that the mechanism of action of the Bxb1 RDF, both in stimulating excision and in inhibiting integration, is mediated by direct interactions in ϕRv1 gpInt or gpInt–DNA complexes (Bibb *et al.*, 2005).

3. Integration specificities of mycobacteriophage integrases

For most phages that encode a tyrosine-integrase, a putative *attP* core site can be identified bioinformatically. The basis for this is the observation that most of these utilize a host tRNA gene for integration, with strand exchange occurring somewhere within the gene, and the phage genome carries the 3′ part of the tRNA gene such that a functional gene is reconstructed following integration. Although recombination itself likely only requires identity between *attP* and *attB* at the 7–8 bp constituting the overlap region between sites of strand cleavage, the requirement for tRNA reconstruction usually extends the sequence identity (or near-identity) to as much as 45 bp, which can be identified readily in a BLASTN search. Furthermore, the *attP* site is typically located near the integrase gene and is usually in an intergenic noncoding interval. As a result, these regions can be used to search sequence databases, followed by determination of whether any matching sequences overlap host tRNA genes. Within the phage genome, it is often possible to identify pairs of short (10–11 bp) sequences flanking the *attP* core that correspond to putative arm-type integrase-binding sites (Morris *et al.*, 2008). Using this strategy, putative *attB* sites can be predicted for most of the mycobacteriophages that encode tyrosine-integrases (Table II). For some of these, including L5, Tweety, BPs, Ms6, and Giles (Freitas-Vieira *et al.*, 1998; Lee *et al.*, 1991; Morris *et al.*, 2008; Pham *et al.*, 2007; Sampson *et al.*, 2009), good experimental evidence supports *attB* site usage. Others await experimental verification. However, for Cluster E phages, as well as LeBron, bioinformatic identification has proven difficult, and *attB* site identification will likely require experimental approaches. One plausible explanation for this is if they either do not use tRNAs for integration or the positions of strand exchange are so close to the 3′ end of a tRNA gene that they are carrying only a minimal segment of homology to the host genome. An alternative explanation is that these phages do not normally infect *M. smegmatis* or any closely related strains, and the *attB* site is simply not present in *M. smegmatis*. This would seem unlikely for Cluster E phages because at least for some, lysogens have been recovered.

Author's personal copy

TABLE II Integration specificities of mycobacteriophage integration sites in *M. smegmatis* mc^{2}155 and *M. tuberculosis* H37Rv

attB	tRNA	M. smeg	M. tb H37Rv	Phages	Cluster	Int
attB-1	tRNA-gly	Msmeg_4676 (4764493–4764563)	NT02MT2675 (2765539–2765609)	L5_33, D29_33, Che12_36, Pukovnik_35	A2	Tyr
				Eagle_32,	A4	Tyr
attB-2	tRNA-Lys	Msmeg_4746 (484790–4847983)	NT02MT2737 (2835492–2835564)	Che8_46, Boomer_46, Llij_40, PMC_38, Tweety_43, Pacc40_40, Ramsey_44	F1	Tyr
attB-3		Msmeg_5156		Bxz2_34	A3	Ser
attB-4	tRNA-Lys	Msmeg_5758 (5834573–5834645)	NT02MT0910 (92387–923798)	Unpublished phages[a]	N	Tyr
attB-5	tRNA-Thr	Msmeg_6152 (6221063–6220991)	NT02MT3969 (4081434–4081359)	Brujita_33, Island3_33	I1	Tyr
attB-6	tRNA-Arg	Msmeg_6349 (6410438–6410366)	NT02MT4110 (4216934–4216862)	BPs_32, Halo_32, Angel_32, Hope_32	G	Tyr
attB-7	GroEL1	MSMEG_0880		Bxb1_35, U2_36, Bethlehem_36 DD5_38, Jasper_38, KBG_38. Lockley_38, Solon_37, SkiPole_40,	A1	Ser
attB-8	tRNA-Tyr	Msmeg_1166 (1228393–1228478)	No	Che9c_41	I2	Tyr
attB-9	tmRNA	Msmeg_2093 (2169257– 2169625)	No	Angelica_41, CrimD_41	K1	Tyr

Author's personal copy

attB-10	tRNA-Ala	MSMEG_2138 (2213142–2213214)	NT02MT3342 (3431909–3431837)	Fruitloop_40, Ardmore_36, Ms6_int	F1	Tyr
attB-11	tRNA-Leu	Msmeg_3245 (3328766–3328690)	NT02MT1769 (1828086–1828010)	Omega_85	J	Tyr
attB-12	tRNA-Pro	Msmeg_3734 (3800622–3800546)	NT02MT1869 (1946611–1946684)	Giles_29	Sin	Tyr
attB-13	tRNA-Met	Msmeg_4452 (4532894–4532821)	NT02MT2502 (2581835–2581762)	Che9d_50	F2	Tyr
Unassigned[b]				Cjw1_53, 244_53, Kostya_53, Porky_51, Pumpkin_54,	E	Tyr
				Peaches_33	A4	Ser
				LeBron_36	L1	Tyr

[a] Two phages have been identified that utilize this site but the phage sequences are as yet incomplete and are not yet published.
[b] *attB* sites have yet to be identified for these phages.

Author's personal copy

Confident bioinformatic identification of *attB* sites for phages encoding serine-integrases is currently not possible. These generally do not integrate into tRNA genes, and the segment of *attP* homology to the host chromosome can be as small as 3 bp (Smith and Thorpe, 2002). The *attB* sites for both Bxb1 (which is likely used for all Subcluster A1 phages because their integrases are extremely similar) and Bxz2 have been identified (Kim *et al*., 2003; Pham *et al*., 2007). The Bxb1 *attB* site is located within the host *groEL1* gene, and integration results in inactivation of the gene with interesting physiological consequences (Kim *et al*., 2003). Specifically, Bxb1 lysogens are defective in the formation of mature biofilms, revealing the novel function of GroEL1, which acts as a dedicated chaperone for mycolic acid biosynthesis (Ojha *et al*., 2005). Bxz2 integrates into the extreme 5′ end of the *M. smegmatis* gene Msmeg_5156 (Pham *et al*., 2007), although no physiological consequences have been examined. The *attB* site for the more distantly related Peaches has yet to be identified. The propensity for phages encoding serine-integrases to integrate within host protein-coding genes with opportunities to influence their physiology makes this class of mycobacteriophages of particular interest.

In total, 13 distinct *attB* sites have been identified or predicted (Table II). To facilitate discussion of these *attB* sites, we have designated them *attB1*–*attB13*, with *attB1* denoting the L5 site. The others are ordered according to their location in the *M. smegmatis* genome, proceeding in a clockwise direction (Fig. 21). Related sites for 9 of these are also present in *M. tuberculosis*; these are numbered according to the same designations (Table II). The placement of these on a circular representation of the *M. tuberculosis* genome illustrates the lack of synteny between these two strains of mycobacteria (Fig. 21). Distribution of *attB* sites is of interest in part because of the utility of using phage integrase-based integration-proficient plasmid vectors, which have the advantage of constructing single-copy recombinants that are genetically stable in the absence of selection (see Section VII.A.1). While additional *attB* specificities would likely be welcome, the current distribution of sites enables the potential use of integration-proficient vectors to introduce genetic elements of choice at different locations relative to the chromosomal origin of DNA replication.

VI. MYCOBACTERIOPHAGE FUNCTIONS ASSOCIATED WITH LYTIC GROWTH

A. Adsorption and DNA injection

Unfortunately, rather little is known about the repertoire of bacterial surface molecules used by mycobacteriophages to specifically recognize their hosts. A *M. smegmatis* peptidoglycolipid, mycoside C(sm), has been

Author's personal copy

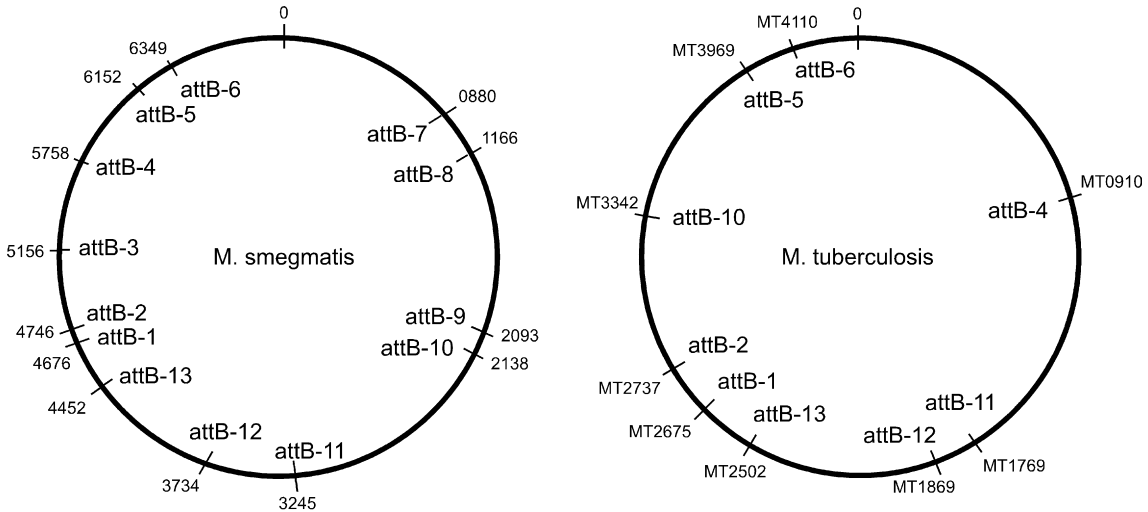


FIGURE 21 Locations of predicted mycobacteriophage *attB* sites in *M. smegmatis* and *M. tuberculosis* genomes. The predicted *attB* sites for all mycobacteriophage genomes containing integrase genes and for which *attP* and *attB* sites can be predicted or are identified experimentally are shown on a circular representation of the *M. smegmatis* genome. Those *attB* sites that are also present in *M. tuberculosis* H37Rv are shown on a circular representation of the *M. tuberculosis* H37Rv genome on the right. The *attB* designation is conserved between the two strains, such that for example *attB-4* has the same sequence in both strains, notwithstanding the different chromosomal position. All *attB* sites and their specific locations are listed in Table II.

Author's personal copy

purified and proposed to play a role in binding of the uncharacterized phage D4 (Furuchi and Tokunaga, 1972), and a set of lyxose-containing molecules have been proposed as receptors for the uncharacterized phage Phlei (Bisso *et al*., 1976; Khoo *et al*., 1996). In addition, a single methylated rhamnose residue on the cell wall-associated glycopeptidolipid has been implicated in the adsorption of phage I3 to *M. smegmatis* (Chen *et al*., 2009). No protein-based receptors for mycobacteriophages have been reported.

Overexpression of a single *M. smegmatis* protein, Mpr, is sufficient to confer high levels of resistance to phage D29 (Barsom and Hatfull, 1996), which may occur through placement of the gene on an extrachromosomal plasmid (Barsom and Hatfull, 1996), expressing it from a strong promoter (Barsom and Hatfull, 1996), or through activation by adjacent transposon insertion (Rubin *et al*., 1999). It is plausible that spontaneous D29 resistance could occur by localized genome amplification of the *mpr* locus that is genetically unstable and recombines back to a single copy in the absence of selection. The cellular role of Mpr is not known and is not present in *M. tuberculosis*. Interestingly, the 215 residue Mpr protein contains a 125 reside domain at its extreme C-terminus belonging to the Telomeric repeat-binding factor 2 (TRF2) superfamily implicated in recognition and binding to TTAGGG-like telomeric repeats. This is an unexpected function for a bacterium that does not contain a linear genome, although it is of interest because of the observation that overexpression of Mpr appears to specifically inhibit injection of D29 DNA. Perhaps this gene has been acquired specifically to prevent phage infection under certain circumstances.

Presumably, specific recognition of the bacterial host is accomplished through structures encoded at the tips of phage tails. Genomic analysis shows that those phages with a siphoviral morphotype encode five to eight putative minor tail protein genes downstream of the tapemeasure protein gene, and for a few phages these have been confirmed experimentally (Ford *et al*., 1998b; Hatfull and Sarkis, 1993; Morris *et al*., 2008). Many mycobacteriophages also encode three to four relatively small genes at the end of the virion structural and assembly operon that may also be involved in tail assembly. Which of these tail proteins is specifically involved in host recognition is unclear. Interestingly, a number of genomes (e.g., in Clusters/Subclusters A1, C1, D, E, F, H1, I1, and J and singletons Corndog and Wildcat) encode a putative β-lactamase-like D-alanyl-D-alanine carboxypeptidase activity that is presumably involved in modification of the cell wall and perhaps facilitates productive association of the tail tip with the cell wall or membrane. However, none of these have been characterized. The lengths of mycobacteriophage tails, especially those with siphoviral morphologies, vary considerably, and the lengths of tapemeasure protein genes vary correspondingly (Pedulla *et al*., 2003). These proteins are of

Author's personal copy

particular interest because many of them encode short sequence motifs associated with peptidoglycan hydrolysis, suggesting functionalities in addition to a role in phage tail assembly (Lai *et al.*, 2006; Pedulla *et al.*, 2003). Initially, three distinct motifs were discovered (Pedulla *et al.*, 2003), although the expanded genomic set suggests that there are at least seven different motifs (L. Marinelli and Graham F. Hatfull, manuscript in preparation). The motif present in TM4 (motif 3) has been shown to be nonessential for viability, although mutants in whom it has been deleted or inactivated have reduced abilities to infect stationary phase cells (Piuri and Hatfull, 2006). Because removal of the motif results in a corresponding reduction in tail length, the tapemeasure protein must be able to adopt two different conformations, an extended rod-like structure involved in tail assembly and a component of the complete tail structure, and a folded structure that has enzymatic activity (Piuri and Hatfull, 2006). Even though peptidoglycan-hydrolyzing motifs cannot be identified in all mycobacteriophages, it seems probable that all can form this proposed alternative folded state. Because most of the mycobacteriophage tapemeasure proteins also contain putative transmembrane-spanning domains—as many as nine in the Cluster G genomes—an intriguing role for all these tapemeasure proteins is as a membrane-located pore through which DNA traverses to gain entry into the cell (L. Marinelli and Graham F. Hatfull, manuscript in preparation).

B. Genome recircularization

Virion DNA is expected to be linear in all mycobacteriophages, and in Clusters A, E, F, G, I, J, K, and L and in singletons Corndog, Giles, and Wildcat, the genomes have defined ends with short ssDNA extension varying from 4 to 14 bases in length; all have 3′ extensions (Table I). Following DNA injection, all are expected to be circularized at an early stage, prior to either DNA replication or integration. The specific requirement for circularization has not been examined, but it is expected that for most phages it is dependent on the action of the host DNA ligase. The kinetics of recirularization are not known.

Two phages with defined ends appear to use a different and an unusual mechanism involving nonhomologous end joining (NJEJ). Phages Omega and Corndog both have 4-base ssDNA extensions, substantially shorter than all the others (9–14 bases), and are different from all other mycobacteriophages in that they also encode a Ku-like protein that facilitates DNA end association in bacterial NHEJ systems (see Sections III.J and III.M.1)(Pitcher *et al.*, 2006). Both *M. smegmatis* and *M. tuberculosis* encode NHEJ systems, including a Ku-like protein and an associated DNA ligase (Lig IV)(Aniukwu *et al.*, 2008; Pitcher *et al.*, 2005). Interestingly, efficient infection of *M. smegmatis* by Corndog and Omega is

Author's personal copy

dependent on the host Lig IV, but not on the host Ku-70 protein (Pitcher *et al.*, 2006). Presumably, the phage-encoded Ku-70 is required for infection, although this has yet to be demonstrated. A conundrum in the implication of NHEJ in genome circularization is that either the phage Ku-70 gene would need to be expressed immediately upon DNA injection or the protein would need to be encapsulated in the phage capsids. Unfortunately, mutants of Omega or Corndog lacking Ku-like genes have yet to be constructed.

It is likely that mycobacteriophages with terminally redundant ends are circularized by homologous recombination. However, it is unclear whether this is dependent on phage-encoded functions or whether host recombinases are utilized. In phage P22 it is proposed that the Erf recombinase, which is essential for phage growth, promotes genome circularization (Botstein and Matz, 1970). We note that mycobacteriophages in Clusters E and L and the singleton Wildcat all encode Erf-family proteins, but their genomes have defined ends, not terminally redundant ends; presumably the Erf-like proteins they encode perform alternative functions. Of those phages that do have terminally redundant ends, only Cluster C phages encode an obvious recombinase, a RecA-like protein (gp201). Cluster B, D, and H phages do not encode a recognizable recombinase at all, thus presumably either exclusively use host recombination enzymes for genome circularization or encode novel recombinases currently uncharacterized.

C. DNA replication

Mycobacteriophage DNA replication represents another understudied but interesting aspect of their biology. Presumably, replication involves a combination of phage- and host-encoded functions and is initiated at one or more origins of replication in the phage genome, although none have been identified. Many of the genomes do not encode their own DNA polymerase and presumably use one or more of the resident polymerases. Others do encode their own DNA polymerase, although both DNA Pol I and Pol III subunits are well represented; LeBron unusually encodes a DNA Pol II-like protein. It has not been shown, however, that any of these are essential for replication or whether host enzymes can be utilized if phage polymerases are inactivated. For the most part, other components of the replication machinery presumably are provided by the host, although we note that Corndog unusually encodes a Pol III clamp loader-like protein (see Section III.M.1). Many genomes also encode predicted DNA primases, although there is great diversity among the types of proteins encoded. For example, in some phages (e.g., in Cluster A), primase functions are associated with two adjacent open reading frames, raising the possibility that a functional enzyme is generated by an

Author's personal copy

unusual translation event (such as a programmed frameshift or a ribosome hop) or by processing at the RNA level. In other phages, the primase function is associated with proteins that also provide either predicted helicase activities (as in Cluster B and K phages) or polymerase functions (as in Clusters D and H and Corndog). Many of the phages also encode apparent stand-alone canonical DNA helicases, frequently of the DnaB family. The gp65 protein of D29 has been characterized and shown to be a structure-specific nuclease with a preference for forked DNA structures (Giri *et al.*, 2009).

Most mycobacteriophages encode a Holliday Junction resolvase, although many different types are represented, including those related to Endo VII, RuvC, and RusA, and they are present in a multitude of genomic locations. Notable exceptions are phages in Clusters D, E, F, and H, raising the possibilities that either these employ different strategies in replication and recombination or encode one or more novel classes of HJ resolvases that have not been recognized previously. Because Holliday junctions are strongly deleterious to DNA packaging and many of the phages encode recombination-promoting proteins, we favor the second of these explanations.

D. Virion assembly

Identification of genes involved in virion structure and assembly is facilitated by their conserved gene order (at least in Siphoviridae), even though the sequences are highly diverse. For example, capsid subunits can be identified in most of the genomes—with perhaps the greatest ambiguity in Cluster H phages (Fig. 11)—although they represent many different sequence phamilies (Hatfull *et al.*, 2010). Nonetheless, it is plausible that they contain similar protein folds to that of the HK97 capsid subunit that is also present in other viruses that lack substantial sequence similarity with it (Fokine *et al.*, 2005; Hendrix, 2005; Johnson, 2010; Wikoff *et al.*, 2000). In some of the genomes there is an identifiable scaffold protein encoded immediately upstream of the capsid gene, and in L5, gp16 has been shown to be a component of head-like particles but absent from intact virions (Hatfull and Sarkis, 1993). Putative scaffold genes are present in many other mycobacteriophage genomes but are divergent at the sequence level and their assignments remain tentative until there is further experimental support. In some mycobacteriophages, the scaffold function may be provided by an N-terminal domain of the capsid protein itself, as in HK97 (Duda *et al.*, 1995). In L5, D29, and TM4 there is strong evidence that the major capsid subunit is covalently cross-linked (Ford *et al.*, 1998a,b; Hatfull and Sarkis, 1993), as described for HK97, and it may be a common feature among mycobacteriophages (Hatfull and Sarkis,

Author's personal copy

1993). But not all do so, and there is evidence against it in Giles (Morris *et al.*, 2008).

As noted earlier, all three phages in Cluster I and the singleton Corndog have prolate heads in contrast to all other mycobacteriophages, which have isometric heads. The dimensions are somewhat different, with Cluster I phages having a length:width ratio of 2.5:1 and Corndog a ratio of 4:1. However, no evident sequence similarity exists between their capsid subunits, although we note that a closely related protein to Cluster I capsid subunits is encoded within the genome of *Streptomyces scabies* (55% amino acid identity), suggesting the presence of a prophage capable of generating prolate-headed particles. It is unclear how the length of these prolate capsids is determined, and at least in Cluster I phages, there is no evidence from genome analysis of genes encoding additional capsid-associated proteins (see Section III.I and Fig. 12). In Corndog, this is less clear because of the greater complexity of the virion structure and assembly operon (Fig. 16).

Like many other phages with siphoviral morphologies, most mycobacteriophages contain a set of four to eight genes located between the major capsid and the major tail subunit that are likely involved in the head–tail joining process. For the most part, these genes are shared among genomes within a subcluster, but only in a few instances are relatives observed in other mycobacteriophages. When they do, they are typically (although not always) as groups of genes that appear to be traveling together; one example is PBI1 genes *20–22* (Fig. 7), which have homologues in similar genomic locations in Cluster H phages (Fig. 11). These observations are consistent with the idea that these head–tail connector proteins have to interact with each other physically.

One of the most highly conserved features of tail assembly genes of Siphoviridae is the expression of two genes between the major tail subunit and the tapemeasure protein genes that are expressed via a programmed translational frameshift to produce tail assembly chaperones (Xu *et al.*, 2004). A programmed frameshift can be identified in nearly all mycobacteriophage genomes, and the majority (Cluster A, C, D, E, and G and Wildcat) use a canonical -1 frameshift as described for phage λ proteins G and G-T (Levin *et al.*, 1993), whereas others (Clusters F and I, Corndog, and Omega) use a +1 (or possibly −2). Somewhat surprising given the strong conservation of this feature among phages of diverse bacterial hosts, no similar frameshifting events have been identified in Cluster B phages.

Some mycobacteriophages have the unusual feature of sharing short but related C-terminal extensions on the ends of their capsid and major tail subunits (Hatfull, 2006). This was first evident from sequencing of the Bxb1 genome (Mediavilla *et al.*, 2000), as the predicted capsid and major tail subunits are both ~85 residues longer than their counterparts in L5 as a consequence of C-terminal extensions. Moreover, Bxb1 extensions are

Author's personal copy

related to each other (47% amino acid identity). Related sequences are present in all other Subcluster A1, A4, B2, and B3 phages situated similarly at the C-terminal ends of their capsids and major tail subunits. HHPred analysis shows predicted structural similarity of these to the C-terminal part of the phage λ major tail subunit (gpV), which is part of the large Big-2 family of Ig-like domains (Fraser *et al.*, 2006; Pell *et al.*, 2009, 2010). Related sequences are found in some other mycobacteriophage proteins, including several copies in a putative minor tail protein in Bxb1 (gp23) and in putative structural proteins in Cluster C1 phages (e.g., Bxz1 gp24). Wildcat and LeBron capsid and major tail subunits have similar types of extensions, and although the sequences are distinct from the others, they are related to other Ig-like domains. The presence of these Ig-like domains is relatively common in phage structural proteins, although their functional roles are unclear. Removal of the C-terminal domain from Lambda gpV results in a 100-fold reduction in viability and a possible defect in tail assembly (Pell *et al.*, 2010), although the relationships between the closely related mycobacteriophage capsids and major tail subunits containing these extensions (e.g., Bxb1) and those that do not (e.g. L5) suggest that these are likely not essential in these mycobacteriophage contexts.

E. Lysis

All mycobacteriophage genomes sequenced to date contain an identifiable lysin A (endolysin) gene encoding a peptidoglycan-hydrolyzing enzyme. However, sequence comparisons show that they are highly modular in nature and encompass a broad span of predicted enzyme specificities; there is no single amino acid sequence motif in common to all. Despite their very different sequences, the Lysin A proteins of D29, Ms6, and TM4 have all been shown to have peptidoglycan-hydrolyzing activity (Garcia *et al.*, 2002; Henry *et al.*, 2010a; Payne *et al.*, 2009). Inactivation of the lysin A gene in Giles (*31*) results in the loss of phage release without interruption of particle assembly (Marinelli *et al.*, 2008; Payne *et al.*, 2009). Delivery of the peptidoglycan hydrolase to its target is likely facilitated by a holin protein; in most mycobacteriophages, a putative holin gene can be identified closely linked to lysin A and encoding a small protein with one or more strongly predicted membrane-spanning domains. However, these are highly diverse at the sequence level and none have been examined experimentally. Interestingly, in phage Ms6, gene *1* product (a close relative of Fruitloop gp28, 97% amino acid identity; Fig. 9) has chaperone-like features and interacts directly with the endolysin to facilitate delivery to its peptidoglycan target in a holin-independent manner (Catalao *et al.*, 2010). Relatives of this protein are also encoded in Subcluster A1 genomes, where it is also closely linked to the lysis genes—and in the Subcluster K1 phage TM4 (gp90), where it is not. Mycobacteriophages are unusual in encoding

Author's personal copy

a second lysis protein, Lysin B, that promotes efficient lysis of the host (Gil *et al.*, 2008; Payne *et al.*, 2009). Deletion of the Giles lysin B gene (*32*) does not lead to loss of viability but gives a reduction of plaque size and the number of particles contained therein (Payne *et al.*, 2009). A few phages do not encode Lysin B, including Che12, Subcluster B2 phages, and the C2 phage Myrna, although not all of these form noticeably small plaques. In Myrna, there is an additional unrelated gene (*244*) implicated in lysis, although it is not known if it substitutes for lysin B activity (Payne *et al.*, 2009). Lysin B has been shown to be a lipolytic enzyme (Gil *et al.*, 2008; Payne *et al.*, 2009), and the crystal structure of D29 Lysin B reveals structural similarity to cutinase family enzymes (Payne *et al.*, 2009). D29 Lysin B has activity as a mycolylarabinogalactan esterase and is proposed to separate the mycolic acid-rich mycobacterial outer membrane from its arabinogalactan anchor (Payne *et al.*, 2009). Ms6 Lysin B acts similarily (Gil *et al.*, 2010). Lysin B can thus be thought of as providing a function analogous to the Rz/Rz1 or spanning proteins encoded by phages of Gram-negative bacteria, which play a role in compromising the integrity of the outer membrane through fusion to the cytoplasmic membrane and facilitating complete lysis (Berry *et al.*, 2008).

VII. GENETIC AND CLINICAL APPLICATIONS OF MYCOBACTERIOPHAGES

Mycobacteriophages have played a central role in the development of tuberculosis genetic systems (Jacobs, 2000), and the large set of sequenced mycobacteriophage genomes provides a rich source of materials for applications both in mycobacterial genetics and in potential clinical applications. Some of these take advantage of the use of whole phage particles, and in these applications the host range is likely to be especially important. Because only Subcluster A2 and Cluster K phages infect *M. tuberculosis* efficiently, these have proven the most useful for these utilities. For development of genetic tools, host range is of less concern because most, if not all of, the genetic functionalities are likely to function equally well in both *M. smegmatis* and *M. tuberculosis*. In at least one example, the reason for lack of phage infection of *M. tuberculosis* can be ascribed to a failure of either adsorption or DNA injection, not phage metabolism per se (R. Dedrick and Graham F. Hatfull, unpublished observations).

A. Genetic tools

1. Integration-proficient vectors

Integration-proficient vectors are those that carry the integration apparatus of a temperate phage and have no other means of DNA replication. The first to be constructed were derived from L5 (Lee *et al.*, 1991),

Author's personal copy

although others with different chromosomal targets have since been reported (Freitas-Vieira *et al.*, 1998; Morris *et al.*, 2008; Murry *et al.*, 2005; Pham *et al.*, 2007). The only phage requirements are the integrase gene and a functional *attP* site; because the *attP* site is typically closely linked to the integrase gene, simple versions of these vectors often can be constructed by inserting a single DNA fragment into a nonreplicating plasmid vector. It should be noted though that although integrase genes can usually be identified readily easily, identification of a functional *attP* is more error-prone. As discussed previously, the core region within which recombination occurs can usually be identified readily, but *attP* function usually requires flanking sequences containing arm-type integrase-binding sites (see Section IV.B.1). In some phages, these too can be predicted bioinformatically (Morris *et al.*, 2008), but in other genomes, this is more difficult. Nonetheless, the functional requirements for *attP* are usually encompassed with a region no larger than about 250–300 bp. Plasmid derivatives can also be used in which the *attP* site and the integrase gene are introduced on separate fragments (Huff *et al.*, 2010).

Of the potential 13 different *attB* sites that can be used for vector construction (Table II, Fig. 21), vectors have been described for at least six of them: *attB-1* (Lee *et al.*, 1991), *attB-2* (Pham *et al.*, 2007), *attB-6* (Sampson *et al.*, 2009), *attB-7* (Kim *et al.*, 2003), *attB-10* (Freitas-Vieira *et al.*, 1998), and *attB-12* (Morris *et al.*, 2008). Two additional site specificities have been described that use integration systems derived from Streptomyces phages or plasmids. Vectors derived from plasmid pSAM2 integrate site specifically into at *attB* overlapping $tRNA^{pro}$ gene Msmeg_6204 and the $tRNA^{pro}$ gene located between Rv3684 and Rv3685c in *M. tuberculosis* H37Rv (Martin *et al.*, 1991; Seoane *et al.*, 1997). Phage phiC31-derived vectors (using a serine-integrase) integrate into an *attB* site located within the putative glutamyl-tRNA(Gln) amidotransferase gene Msmeg_3400 and presumably inactivate it; there are three potential *attB* sites in *M. tuberculosis* (Murry *et al.*, 2005). L5 integration-proficient plasmids have also been manipulated such as to carry an additional *attB* site that will accept secondary integration events (Saviola and Bishai, 2004), and the Ms6 system has been manipulated so as to use alternative $tRNA^{ala}$ genes as integration sites (Vultos *et al.*, 2006). Integrated sequences can also be switched efficiently by introduction of a second plasmid with the same integration specificity and a second selectable marker (Pashley and Parish, 2003).

Integration-proficient plasmid vectors have several advantages over extrachromosomally replicating vectors. For example, introduction of genes in single copy typically avoids the overexpression seen with extrachromosomal vectors and thus avoids complications that can be encountered in complementation experiments. Second, they can have greater stability in the absence of selection relative to extrachromosomal vectors,

Author's personal copy

provided that the phage-encoded excise functions are not also present. Even in the absence of excise, integrase-mediated excise-independent excisive recombination can lead to plasmid loss, which may be exacerbated in recombinants that are at a selective disadvantage (Springer *et al.*, 2001). Improved stability can be provided if the integrase gene is absent from the recombinant, which can be accomplished either by introducing the integrase gene on a second, nonreplicating plasmid that is subsequently lost (Peña *et al.*, 1997) or by site specifically removing the integrase gene from the recombinant (Huff *et al.*, 2010). Phage-encoded, site-specific recombination systems are also useful for efficient modification of recombinants, and excisive recombination by L5 integrase has been used to demonstrate gene essentiality (Parish *et al.*, 2001).

2. Selection by immunity

Temperate phages are immune to phage superinfection. If the superinfecting phage is defective in lysogeny and thus efficiently kills the bacterial cells, then this provides an effective means for using phage immunity functions—and repressor genes specifically—as selectable markers. This has been demonstrated using the L5 repressor gene (*71*) and using either D29 or a clear-plaque mutant of L5 (Donnelly-Wu *et al.*, 1993). Phage particles can be spread readily onto solid media prior to plating of cells, and relatively large numbers of cells can be plated and still get efficient killing of nontransformed cells. The obvious advantage of such systems is that they avoid the use of antibiotics and are thus useful for constructing complex recombinants where relatively few markers are available, for manipulation of strains that are extensively drug resistant even without manipulation, and to minimize biosafety concerns of generating highly drug-resistant forms of pathogenic strains. Because there are a substantial number of distinct mycobacteriophage immune specificities (see Section V.A), there is the potential to construct a large collection of compatible markers, at least for *M. smegmatis*. Although there are a large number of phages that encode identifiable integrases, only in Cluster A and Cluster G phages have repressor genes been identified. Phage repressors appear to be highly diverse and, in many cases, will need to be identified experimentally. We also note that immune selection requires the isolation of a clear-plaque derivative of the phage, and we note that because many of the temperate mycobacteriophages often form lysogens at relatively low frequencies, this is not always a simple task.

3. Generalized transduction

Generalized transduction is one of the most useful tools broadly implemented in bacterial genetics because it provides a simple means of moving genetic markers and mutations into different strain backgrounds. As such, it becomes easy to construct isogenic strains and to thus draw

Author's personal copy

confident conclusions about the correlation between genotype and phenotype. Generalized transducing phages are typically those that package their DNA by headful-packaging systems and thus have genomes that are terminally redundant and circularly permuted. As described earlier, there are many mycobacteriophages in this class, including those in Clusters B, C, D, and H.

The first mycobacteriophage demonstrated to mediate generalized transduction of *M. smegmatis* was I3 (Raj and Ramakrishnan, 1970), a myovirus whose complete genome has not been sequenced, but is likely to be a Cluster C-like phage. It has been shown subsequently that Bxz1 is also a generalized transducing phage and can be used to efficiently exchange genetic markers between strains of *M. smegmatis* (Lee *et al.*, 2004). It is highly likely that other members of Cluster C behave similarly. Transduction by phages of Clusters A, D, and H has yet to be demonstrated.

No phages capable of generalized transduction of *M. tuberculosis* have been identified. This is unfortunate because there is a particular need for such phages to construct isogenic strains, especially for the analysis of mutations that occur in clinical isolates and that may be suspected of contributing to drug resistance or pathogenicity phenotypes. We note that none of the known mycobacteriophages with circularly permuted terminally redundant genomes infect *M. tuberculosis*. Although such phages may well exist in nature and await isolation, there is also the possibility that the idiosyncrasies of homologous recombination systems in *M. tuberculosis*—especially their proclivity for illegitimate recombination of linear DNA substrates introduced by electroporation (Kalpana *et al.*, 1991)—could have thwarted the successful evolution of such phages.

4. Transposon delivery

Several transposons have been described that can be used for insertional mutagenesis in *M. tuberculosis* and *M. smegmatis* (Cirillo *et al.*, 1991; Fomukong and Dale, 1993; Guilhot *et al.*, 1992; Rubin *et al.*, 1999). Mycobacteriophages offer attractive systems for transposon delivery because of the high efficiency of infection and the ability to deliver transposon DNA to nearly every cell in a liquid culture. This is especially important given the relatively low frequencies of movement of most transposons in bacteria. Phage delivery of transposons has the additional advantage over plasmid delivery systems in that the mutants recovered result from independent transposition events, providing the opportunity to generate mutant libraries of a large number of different insertions (Lamichhane *et al.*, 2003; Sassetti *et al.*, 2001), which is critical for applications such as transposon site hybridization (Sassetti *et al.*, 2003).

For efficient phage-mediated transposon delivery, it is important that lytic growth of the phage does not lead to cell death (Kleckner *et al.*, 1991).

Author's personal copy

Conditionally replicating mutants of both D29 and TM4 have been described that grow normally at 30°C but fail to replicate at higher temperatures (37°C for TM4, 38.5°C for D29)(Bardarov *et al.*, 1997). These mutants were isolated to ensure low frequencies of reversion to wild-type replication patterns, a potential concern when seeking selection of relatively low frequency transposition events. Coupling of these mutant phages with shuttle phasmids enables introduction of a variety of transposons of choice and has created a facile system for mutagenesis of a variety of mycobacterial strains. The phasmids can be prepared and grown in *M. smegmatis* at 30°C, but then used to infect *M. smegmatis* or *M. tuberculosis* at the nonpermissive temperature and selection for transposon mutants on solid media (Bardarov *et al.*, 1997).

5. Specialized transducing phages

Conditionally replicating mycobacteriophages also provide a powerful approach to the delivery of allelic exchange substrates for constructing mycobacterial mutants, including gene knockout and gene replacement mutants (Bardarov *et al.*, 2002). The approach is similar to that for transposon delivery, and construction of a phasmid carrying a DNA substrate in which an antibiotic resistance marker is flanked by 500–1000 bp corresponding to the flanking sequences of the gene to be replaced. Following infection, gene replacement mutants can be selected by antibiotic resistance. Because effectively every cell can be infected by the phage, the number of recombinants should be very high, even if gene replacement occurs in only a relatively small proportion of cells. In practice, only perhaps 10^{-6} or fewer cells generate recombinants, although a very high proportion of these result from homologous recombination at the intended site (Bardarov *et al.*, 1997), in contrast to the high proportion of illegitimate events observed when introducing linear DNA fragments by electroporation (Kalpana *et al.*, 1991). The reason why the recovery of recombinants is relatively inefficient is not known, although it suggests that there is a substantial opportunity to increase the recovery of the number of recombinants.

6. Mycobacterial recombineering

Recombineering [genetic engineering using recombination (Court *et al.*, 2002)] offers a general approach to constructing mutant bacterial derivatives by taking advantage of the high frequencies of homologous recombination that can be accomplished by the expression of phage-encoded recombination systems. Perhaps the most widely used system in *E. coli* is the λ-encoded Red system in which three proteins, Exo, Beta, and Gam, contribute to recombination proficiency. Exo is an exonuclease that degrades one strand of dsDNA substrates, Beta is a protein that promotes pairing of complementary DNA strands, and Gam is an

Author's personal copy

inhibitor of RecBCD (Court *et al*., 2002). When either dsDNA or short ssDNA substrates are introduced into *E. coli* by electroporation, recombination with a chromosomal or plasmid target occurs efficiently; in some configurations, desired recombinants can be identified even without selection. Similar systems have been described that utilize the RecET system encoded by the *E. coli rac* prophage (Murphy, 1998; Zhang *et al*., 1998).

The *E. coli* recombineering systems do not function well in mycobacteria, especially when using dsDNA substrates (van Kessel and Hatfull, 2007, 2008a). Mycobacterial-specific recombineering systems have been developed using mycobacteriophage-encoded recombinases, especially those related to the RecET systems (van Kessel and Hatfull, 2007, 2008a, b; van Kessel *et al*., 2008), such as genes *60* and *61* of phage Che9c (Fig. 12). When both Che9c gp60 and gp61 are expressed from an inducible expression system in *M. smegmatis* or *M. tuberculosis*, recombination frequencies are elevated substantially. Introduction of a dsDNA allelic exchange substrate in which 500–1000 bp of chromosomal homology flank an antibiotic resistance marker, followed by selection, generates recombinants efficiently (van Kessel and Hatfull, 2007). dsDNA recombineering works well and reproducibly in *M. smegmatis*, but anecdotal reports suggest that it may be somewhat more erratic in *M. tuberculosis*, perhaps due to irreproducibility of efficient expression of the recombinases.

Recombineering using ssDNA substrate requires only short synthetic oligonucleotide-derived substrates, provided that mutations are introduced that confer a selectable phenotype (van Kessel and Hatfull, 2008a). Interestingly, in both *M. smegmatis* and *M. tuberculosis* there is a very substantial strand bias, such that oligonucleotides with complementary sequences can yield recombinants at frequencies differing by more than 10^4-fold (van Kessel and Hatfull, 2008a). For engineering purposes it is therefore important that the most efficient of the two possible oligonucleotides is used, which is usually that corresponding to the leading strand of chromosomal DNA replication (i.e., can base pair with the template for lagging strand synthesis). ssDNA recombineering can be used to generate recombinants in the absence of direct selection using coelectroporation of two oligonucleotides: one designed to introduce the desired mutation and one that can be used for selection. A high proportion of selected recombinants also carry the unselected mutation and can be detected by physical screening (van Kessel and Hatfull, 2008a).

Recombineering provides an especially powerful tool for genetic manipulation of the mycobacteriophages themselves (Marinelli *et al*., 2008; van Kessel *et al*., 2008). The Bacteriophage Recombineering of Electroporated DNA (BRED) system involves coelectroporation of a phage genomic DNA substrate and a short (~200 bp) dsDNA substrate in a strain in which recombineering functions have been induced. Plaques

Author's personal copy

can then be recovered on solid media in an infectious center configuration in which each electroporated cell that has taken up phage DNA gives rise to a plaque. When individual plaques are screened for the presence of either wild-type or mutant alleles at the targeted site, all contain the wild-type allele, but 10% or more also contain the mutant allele (Marinelli *et al.*, 2008). The desired phage mutant can then be recovered from this mixed primary plaque by replating and testing individual secondary plaques. In this way, two rounds of polymerase chain reaction analysis of 12–18 plaques typically generates the desired mutant, provided that the mutant is viable. In at least some cases, nonviable plaques can be recovered by complementation (Marinelli *et al.*, 2008; Payne *et al.*, 2009). BRED can be used to introduce insertions, deletions, and point mutations into mycobacteriophage genomes (Marinelli *et al.*, 2008).

B. Clinical tools

1. Phage-based diagnosis of *M. tuberculosis*

The ability of mycobacteriophages to infect mycobacterial hosts specifically and efficiently has led to three types of systems for the diagnosis of *M. tuberculosis* infections. There is a particular need for such systems because the diagnosis of human tuberculosis is complicated by the slow growth of the bacteria, the need to determine drug susceptibility profiles, and the fact that the demographic and geographic areas of greatest need often have only minimal resources to devote to this issue. An inexpensive, rapid, simple diagnostic system for drug susceptibility testing of *M. tuberculosis* is therefore highly desirable.

The first phage-based diagnostic developed was the phage-typing approach in which a substantial number of mycobacteriophages were isolated whose host ranges were informative about the identity of any unknown host (Engel, 1975; Redmond and Ward, 1966). In this way, an unknown clinical isolate could be tested for susceptibility to a set of phages and preliminary identification was obtained within a few days. Of particular note in this regard is the use of phage DS6A, whose host range is restricted to bacteria of the *M. tuberculosis* complex, including *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium canetti*, and *Mycobacterium microti* (Bowman, 1969; Jones, 1975). DS6A has not yet been characterized genomically. Although phage typing is useful for strain identification, it does not readily provide information about drug susceptibility profiles.

A second phage-based diagnostic system is the phage amplification biological assay (PhaB), which is based on the ability of mycobacteriophages to infect and amplify in *M. tuberculosis* if present in a clinical sample, followed by enumeration of particles using *M. smegmatis* as a host (Eltringham *et al.*, 1999; Watterson *et al.*, 1998; Wilson *et al.*, 1997).

Author's personal copy

Phage D29 has been the primary focus for this approach because it infects both *M. tuberculosis* and *M. smegmatis* and produces large, clear, easily identifiable plaques. The system has been evaluated with clinical specimens in several studies and has been used to discern rifampicin-resistant and rifampicin-sensitive hosts (Albert *et al.*, 2001, 2002a,b, 2004; McNerney *et al.*, 2000; Pai *et al.*, 2005). The third approach is the use of reporter mycobacteriophages in which recombinant phages carrying a reporter gene, such as firefly luciferase (FFlux) or GPF (or related fluorescence genes), can be used to detect the physiological status of the cell rapidly, thus reporting on drug susceptibilities (Jacobs *et al.*, 1993; Piuri *et al.*, 2009). These phages can be constructed readily using either shuttle phasmid technology or BRED recombineering (Marinelli *et al.*, 2008) and can be used in several configurations depending on the reporter gene used and the detection technology available. Fluoromycobacteriophages have some notable advantages in that it is possible to detect single cells following infection and the signal is retained after fixation, providing additional biosafety and assay flexibility (Piuri *et al.*, 2009). The assay is rapid, and the use of light-emitting, diode-based microscopes provides a potentially simple clinical configuration. Establishment of efficient phage infection conditions directly in sputum samples remains the highest priority for direct clinical evaluation. Although reporter phages have been derived from TM4 (Jacobs *et al.*, 1993), D29 (Pearson *et al.*, 1996), and L5 (Sarkis *et al.*, 1995), none of these are specific to *M. tuberculosis*, and as with the PhaB assay, use of *M. tuberculosis*-specific phages would be advantageous.

2. Phage therapy?

Mycobacteriophages would seem to have some advantages for direct therapeutic treatment of pulmonary tuberculosis, especially in circumstances in which MDR-TB and XDR-TB infections respond poorly to antibiotic therapy. Delivery directly to the lung would seem feasible, and there are reports of evaluation in animal model systems (Koz'min-Sokolov and Vabilin, 1975; Sula *et al.*, 1981). The potential disadvantage is that the phage particles may not gain access to bacteria that are intracellular, or contained with granulomas, and therefore a therapeutic cure would seem improbable. Bronxmeyer and colleagues (2002) have explored successfully the possibility of using *M. smegmatis* as a surrogate to deliver TM4 to *M. tuberculosis*-infected macrophages, suggesting a novel route to killing intracellular bacteria (Broxmeyer *et al.*, 2002) and circumventing this problem. Phage resistance poses another potential concern, which could potentially be overcome by using either serial applications of phages to which different mechanisms of phage resistance occur or phage cocktails with broad combinations of phages. The actual number of phages currently available for such an application is rather

Author's personal copy

small, with D29 being the most attractive candidate. If phage therapy is to be evaluated, it will be important to identify additional mycobacteriophages that infect both *M. tuberculosis* and *M. smegmatis* (for propagation purposes), kill a very high proportion of bacterial cells upon infection, and represent different patterns of host resistance responses. A related application is the possibility of using mycobacteriophages to interfere with active dissemination of tuberculosis from an actively infected person to household contacts, family members, and/or co-workers. Because dissemination likely involves forms of bacteria susceptible to phage infection, application of a suitable phage preparation by inhalation, aspiration, or nebulization could reduce the number of *M. tuberculosis* cells passing through the upper respiratory tract greatly and reduce the chances of transfer to an uninfected individual. An especially attractive configuration would be to use phages in a prophylactic form to protect those in close contact from acquisition from a patient, while enabling the infected person to undergo a normal course of antibiotic therapy. This would also minimize opportunities for the selection of phage-resistant mutants because the number of bacteria in contact with the phage is relatively small. However, success of this approach is anticipated to depend on the ability to deliver an effective quantity of phage particles, stability of the particles, and the likelihood that multiple doses over a period of time will be required for maximum effectiveness.

VIII. FUTURE DIRECTIONS

As the collection of sequenced mycobacteriophage genomes has grown, it has become abundantly clear how much we really do not understand about this fascinating group of viruses. For the most part, future directions are reasonably clear, and five major paths can be envisaged.

First, it is clear that much more needs to be learned about the genetic diversity of mycobacteriophages. As more mycobacteriophage genomes are sequenced, the numbers of more closely related phages have grown, but entirely new genomes continue to emerge, as well as phages related to those classified previously as singletons. The combination of an immensely powerful and high-impact integrated research and education platform for phage isolation, and the dramatic decline in genomic sequencing costs, will help fuel this ongoing effort in mycobacteriophage genome and discovery. It is not unreasonable to suppose that the collection of sequenced mycobacteriophage genomes could rise to more than 1000 within the next 5 years. Presumably, at some point we will reach the point of genomic saturation where further genomic sequencing will provide diminishing returns, but it is unclear when that will be reached. We note that although isolating entirely new genomes is thrilling, the

Author's personal copy

collection of groups of related phages provides powerful resources for understanding the detailed mechanisms of genome evolution. Current phages all share a common host in *M. smegmatis* mc^2155, but preliminary observations (C. Bowman and Graham F. Hatfull) show that some of these can discriminate between different substrains of *M. smegmatis*. It is thus likely that use of other *M. smegmatis* strains for phage isolation or different mycobacterial species will yet further expand the amazing diversity of the mycobacteriophage population. Moreover, it is critical that additional phages that infect *M. tuberculosis* be isolated using *M. tuberculosis* itself either as a host or as a surrogate that is much more closely related to it than *M. smegmatis*. Second, it will be important to establish the detailed host specificities of the sequenced mycobacteriophages. A plausible reason for their great diversity in nucleotide sequence, genome length, and GC% is that they share different but overlapping host ranges, with *M. smegmatis* mc^2155 being the common host. One approach would be to test the susceptibilities of known strains within Actinomycetales for infection by the mycobacteriophages, although it is important to recognize that these may poorly reflect the full diversity of bacteria in the environments from which the phages are isolated. Another approach would be to characterize the bacterial population of the samples from where phages are isolated more extensively, although this is complicated by the massive complexity of the soil biome and the likelihood that many of them, including potential mycobacteriophage hosts, are not cultivatable.

A third major area of focus should be on determination of what the many unknown mycobacteriophage genes do, using both functional genomic and structural genomic approaches. The BRED engineering technology provides a powerful means of constructing defined phage mutants, including gene knockouts and point mutations, and large numbers of mutants can be generated and characterized readily. Thus it is now possible to apply functional genomic approaches to whole genomes and to dissect them genetically. Structural genomic approaches will also be useful, especially as many of the phage-encoded genes are small, and the encoded proteins should be amenable to structural analysis by crystallography and nuclear magnetic resonance. Structural information should provide clues as to potential functions, but phage proteins may also be a rich source of novel protein folds, especially given their vast sequence diversity. These functional genomic and structural approaches are immensely powerful, although the sheer number of genes to analyze makes this an important but daunting prospect.

The fourth major direction is to characterize the patterns of mycobacteriophage genome expression, identify the signals for transcription initiation and termination, and elucidate the mechanisms of gene regulation. Little is known about the global patterns of mycobacteriophage gene expression, although the genomes should be amenable to transcriptome

Author's personal copy

analysis using either microarrays or high throughput RNA-seq. In only a small number of examples have putative promoters been identified, and it is clear that many promoters cannot be readily identified bioinformatically. Investigation of gene expression and its regulation is expected to be especially rewarding, as previous studies reveal an abundance of novelty, such as with the remarkable stoperator system in Cluster A phages, and new systems for lytic–lysogenic decision systems in Cluster G phages. Moreover, it seems likely that some mycobacteriophage-encoded proteins are expressed from prophages with the capacity to influence host physiology, and a combination of expression and functional studies may provide important clues, especially in examples where phage-encoded proteins may influence pathogenicity. Finally, there are numerous potential routes to exploit the mycobacteriophages to develop additional tools for mycobacterial genetics. These range from additional integration-proficient vectors with novel *attB* target specificities, a suite of repressor-mediated selectable markers, and regulated expression systems to mycobacteriophage-specific packaging systems, mycobacteriophage-based antigen display systems, new tools for mutagenesis, and applications for diagnosis and therapy. The world is your oyster.

ACKNOWLEDGMENTS

I thank all of my colleagues in Pittsburgh for their long-standing and ongoing collaborations, including Roger Hendrix, Jeffrey Lawrence, Craig Peebles, Deborah Jacobs-Sera, Welkin Pope, Dan Russell, Bekah Dedrick, Greg Broussard, Anil Ojha, and Pallavi Ghosh, and the many graduate students and research assistants who have contributed to this work. I also thank Dr. Bill Jacobs and his colleagues at Albert Einstein College of Medicine and my colleagues at the HHMI Science Education Alliance, including Tuajuanda Jordan, Lucia Barker, Kevin Bradley, and Razi Khaja. I am especially grateful to the large number of individual high school and undergraduate phage hunters both at Pittsburgh and in the SEA-PHAGES programs that have contributed broadly to the advancement of our understanding of the mycobacteriophages. I extend special thanks to Dan Russell for help with Figure 2 and the GC% analysis and to Deborah Jacobs-Sera, Welkin Pope, and Roger Hendrix for helpful comments on the manuscript.

REFERENCES

Abedon, S. T. (2009). Phage evolution and ecology. *Adv. Appl. Microbiol.* **67:**1–45.

Albert, H., Heydenrych, A., Brookes, R., Mole, R. J., Harley, B., Subotsky, E., Henry, R., and Azevedo, V. (2002a). Performance of a rapid phage-based test, FASTPlaqueTB, to diagnose pulmonary tuberculosis from sputum specimens in South Africa. *Int. J. Tuberc. Lung Dis.* **6**(6)**:**529–537.

Albert, H., Heydenrych, A., Mole, R., Trollip, A., and Blumberg, L. (2001). Evaluation of FASTPlaqueTB-RIF, a rapid, manual test for the determination of rifampicin resistance from *Mycobacterium tuberculosis* cultures. *Int. J. Tuberc. Lung Dis.* **5**(10)**:**906–911.

Author's personal copy

Albert, H., Trollip, A., Seaman, T., and Mole, R. J. (2004). Simple, phage-based (FASTPplaque) technology to determine rifampicin resistance of *Mycobacterium tuberculosis* directly from sputum. *Int. J. Tuberc. Lung Dis.* **8**(9)**:**1114–1119.

Albert, H., Trollip, A. P., Mole, R. J., Hatch, S. J., and Blumberg, L. (2002b). Rapid indication of multidrug-resistant tuberculosis from liquid cultures using FASTPlaqueTB-RIF, a manual phage-based test. *Int. J. Tuberc. Lung Dis.* **6**(6)**:**523–528.

Aniukwu, J., Glickman, M. S., and Shuman, S. (2008). The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. *Genes Dev.* **22**(4)**:**512–527.

Bandhu, A., Ganguly, T., Chanda, P. K., Das, M., Jana, B., Chakrabarti, G., and Sau, S. (2009). Antagonistic effects Na+ and Mg2+ on the structure, function, and stability of mycobacteriophage L1 repressor. *BMB Rep.* **42**(5)**:**293–298.

Bandhu, A., Ganguly, T., Jana, B., Mondal, R., and Sau, S. (2010). Regions and residues of an asymmetric operator DNA interacting with the monomeric repressor of temperate mycobacteriophage L1. *Biochemistry* **49**(19)**:**4235–4243.

Bardarov, S., Bardarov, S., Jr., Pavelka, M. S., Jr., Sambandamurthy, V., Larsen, M., Tufariello, J., Chan, J., Hatfull, G., and Jacobs, W. R., Jr. (2002). Specialized transduction: An efficient method for generating marked and unmarked targeted gene disruptions in *Mycobacterium tuberculosis*. *M. bovis BCG and M. smegmatis. Microbiology* **148** (Pt 10)**:**3007–3017.

Bardarov, S., Kriakov, J., Carriere, C., Yu, S., Vaamonde, C., McAdam, R. A., Bloom, B. R., Hatfull, G. F., and Jacobs, W. R., Jr. (1997). Conditionally replicating mycobacteriophages: A system for transposon delivery to *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **94**(20)**:**10961–10966.

Barsom, E. K., and Hatfull, G. F. (1996). Characterization of *Mycobacterium smegmatis* gene that confers resistance to phages L5 and D29 when overexpressed. *Mol. Microbiol.* **21**(1)**:**159–170.

Berry, J., Summer, E. J., Struck, D. K., and Young, R. (2008). The final step in the phage infection cycle: The Rz and Rz1 lysis proteins link the inner and outer membranes. *Mol. Microbiol.* **70**(2)**:**341–351.

Bhattacharya, B., Giri, N., Mitra, M., and Gupta, S. K. (2008). Cloning, characterization and expression analysis of nucleotide metabolism-related genes of mycobacteriophage L5. *FEMS Microbiol. Lett.* **280**(1)**:**64–72.

Bibb, L. A., Hancox, M. I., and Hatfull, G. F. (2005). Integration and excision by the large serine recombinase phiRv1 integrase. *Mol. Microbiol.* **55**(6)**:**1896–1910.

Bibb, L. A., and Hatfull, G. F. (2002). Integration and excision of the *Mycobacterium tuberculosis* prophage-like element, phiRv1. *Mol. Microbiol.* **45**(6)**:**1515–1526.

Bisso, G., Castelnuovo, G., Nardelli, M. G., Orefici, G., Arancia, G., Laneelle, G., Asselineau, C., and Asselineau, J. (1976). A study on the receptor for a mycobacteriophage : phage phlei. *Biochimie* **58**(1–2)**:**87–97.

Botstein, D., and Matz, M. J. (1970). A recombination function essential to the growth of bacteriophage P22. *J. Mol. Biol.* **54**(3)**:**417–440.

Bowman, B., Jr. (1958). Quantitative studies on some mycobacterial phage-host systems. *J. Bacteriol.* **76**(1)**:**52–62.

Bowman, B. U. (1969). Properties of mycobacteriophage DS6A. *I. Immunogenicity in rabbits. Proc. Soc. Exp. Biol. Med.* **131**(1)**:**196–200.

Brown, K. L., Sarkis, G. J., Wadsworth, C., and Hatfull, G. F. (1997). Transcriptional silencing by the mycobacteriophage L5 repressor. *EMBO. J.* **16**(19)**:**5914–5921.

Broxmeyer, L., Sosnowska, D., Miltner, E., Chacon, O., Wagner, D., McGarvey, J., Barletta, R. G., and Bermudez, L. E. (2002). Killing of *Mycobacterium avium* and *Mycobacterium tuberculosis* by a mycobacteriophage delivered by a nonvirulent mycobacterium: A model for phage therapy of intracellular bacterial pathogens. *J. Infect. Dis.* **186** (8)**:**1155–1160.

Author's personal copy

Caruso, S. M., Sandoz, J., and Kelsey, J. (2009). Non-STEM undergraduates become enthusiastic phage-hunters. *CBE Life Sci. Educ.* **8**(4)**:**278–282.

Casas, V., and Rohwer, F. (2007). *Phage metagenomics. Methods Enzymol.* **421:**259–268.

Catalao, M. J., Gil, F., Moniz-Pereira, J., and Pimentel, M. (2010). The mycobacteriophage Ms6 encodes a chaperone-like protein involved in the endolysin delivery to the peptidoglycan. *Mol. Microbiol.* **77**(3)**:**672–686.

Chattoraj, P., Ganguly, T., Nandy, R. K., and Sau, S. (2008). Overexpression of a delayed early gene hlg1 of temperate mycobacteriophage L1 is lethal to both *M. smegmatis* and *E. coli*. *BMB Rep.* **41**(5)**:**363–368.

Chen, J., Kriakov, J., Singh, A., Jacobs, W. R., Jr., Besra, G. S., and Bhatt, A. (2009). Defects in glycopeptidolipid biosynthesis confer phage I3 resistance in *Mycobacterium smegmatis*. *Microbiology* **155**(Pt 12)**:**4050–4057.

Cirillo, J. D., Barletta, R. G., Bloom, B. R., and Jacobs, W. R., Jr. (1991). A novel transposon trap for mycobacteria: isolation and characterization of IS1096. *J. Bacteriol.* **173**(24)**:**7772–7780.

Clark, A. J., Inwood, W., Cloutier, T., and Dhillon, T. S. (2001). Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. *J. Mol. Biol.* **311**(4)**:**657–679.

Colangeli, R., Haq, A., Arcus, V. L., Summers, E., Magliozzo, R. S., McBride, A., Mitra, A. K., Radjainia, M., Khajo, A., Jacobs, W. R., Jr., Salgame, P., and Alland, D. (2009). The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates. *Proc. Natl. Acad. Sci. USA* **106**(11)**:**4414–4418.

Colangeli, R., Helb, D., Vilcheze, C., Hazbon, M. H., Lee, C. G., Safi, H., Sayers, B., Sardone, I., Jones, M. B., Fleischmann, R. D., Peterson, S. N., Jacobs, W. R., Jr., *et al.* (2007). Transcriptional regulation of multi-drug tolerance and antibiotic-induced responses by the histone-like protein Lsr2 in *M. tuberculosis*. *PLoS Pathog* **3**(6)**:**e87.

Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H., and Prangishvili, D. (2008). *Exploring the prokaryotic virosphere. Res. Microbiol.* **159**(5)**:**306–313.

Court, D. L., Sawitzke, J. A., and Thomason, L. C. (2002). Genetic engineering using homologous recombination. *Annu. Rev. Genet.* **36:**361–388.

Datta, H. J., Mandal, P., Bhattacharya, R., Das, N., Sau, S., and Mandal, N. C. (2007). The G23 and G25 genes of temperate mycobacteriophage L1 are essential for the transcription of its late genes. *J. Biochem. Mol. Biol.* **40**(2)**:**156–162.

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

Donnelly-Wu, M. K., Jacobs, W. R., Jr., and Hatfull, G. F. (1993). Superinfection immunity of mycobacteriophage L5: Applications for genetic transformation of mycobacteria. *Mol. Microbiol.* **7**(3)**:**407–417.

Duda, R. L., Hempel, J., Michel, H., Shabanowitz, J., Hunt, D., and Hendrix, R. W. (1995). Structural transitions during bacteriophage HK97 head assembly. *J. Mol. Biol.* **247**(4)**:** 618–635.

Eltringham, I. J., Wilson, S. M., and Drobniewski, F. A. (1999). Evaluation of a bacteriophage-based assay (phage amplified biologically assay) as a rapid screen for resistance to isoniazid, ethambutol, streptomycin, pyrazinamide, and ciprofloxacin among clinical isolates of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **37**(11)**:**3528–3532.

Engel, H. W. (1975). Phage typing of strains of ''*M. tuberculosis*'' in the Netherlands *Ann. Sclavo* **17**(4)**:**578–583.

Fineran, P. C., Blower, T. R., Foulds, I. J., Humphreys, D. P., Lilley, K. S., and Salmond, G. P. (2009). The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proc. Natl. Acad. Sci. USA* **106**(3)**:**894–899.

Fokine, A., Leiman, P. G., Shneider, M. M., Ahvazi, B., Boeshans, K. M., Steven, A. C., Black, L. W., Mesyanzhinov, V. V., and Rossmann, M. G. (2005). Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl. Acad. Sci. USA* **102**(20)**:**7163–7168.

Author's personal copy

Fomukong, N. G., and Dale, J. W. (1993). Transpositional activity of IS986 in *Mycobacterium smegmatis*. *Gene* **130**(1)**:**99–105.

Ford, M. E., Sarkis, G. J., Belanger, A. E., Hendrix, R. W., and Hatfull, G. F. (1998a). Genome structure of mycobacteriophage D29: Implications for phage evolution. *J. Mol. Biol.* **279** (1)**:**143–164.

Ford, M. E., Stenstrom, C., Hendrix, R. W., and Hatfull, G. F. (1998b). Mycobacteriophage TM4: Genome structure and gene expression. *Tuber. Lung Dis.* **79**(2)**:**63–73.

Fraser, J. S., Yu, Z., Maxwell, K. L., and Davidson, A. R. (2006). Ig-like domains on bacteriophages: A tale of promiscuity and deceit. *J. Mol. Biol.* **359**(2)**:**496–507.

Freitas-Vieira, A., Anes, E., and Moniz-Pereira, J. (1998). The site-specific recombination locus of mycobacteriophage Ms6 determines DNA integration at the tRNA(Ala) gene of *Mycobacterium* spp. *Microbiology* **144**(Pt 12)**:**3397–3406.

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

Fullner, K. J., and Hatfull, G. F. (1997). Mycobacteriophage L5 infection of *Mycobacterium bovis* BCG: Implications for phage genetics in the slow-growing mycobacteria. *Mol. Microbiol.* **26**(4)**:**755–766.

Furuchi, A., and Tokunaga, T. (1972). Nature of the receptor substance of *Mycobacterium smegmatis* for D4 bacteriophage adsorption. *J. Bacteriol.* **111**(2)**:**404–411.

Ganguly, T., Bandhu, A., Chattoraj, P., Chanda, P. K., Das, M., Mandal, N. C., and Sau, S. (2007). Repressor of temperate mycobacteriophage L1 harbors a stable C-terminal domain and binds to different asymmetric operator DNAs with variable affinity. *Virol. J.* **4:**64.

Ganguly, T., Chanda, P. K., Bandhu, A., Chattoraj, P., Das, M., and Sau, S. (2006). Effects of physical, ionic, and structural factors on the binding of repressor of mycobacteriophage L1 to its cognate operator DNA. *Protein Pept. Lett.* **13**(8)**:**793–798.

Ganguly, T., Chattoraj, P., Das, M., Chanda, P. K., Mandal, N. C., Lee, C. Y., and Sau, S. (2004). A point mutation at the C-terminal half of the repressor of temperate mycobacteriophage L1 affects its binding to the operator DNA. *J. Biochem. Mol. Biol.* **37**(6)**:**709–714.

Garcia, M., Pimentel, M., and Moniz-Pereira, J. (2002). Expression of Mycobacteriophage Ms6 lysis genes is driven by two sigma(70)-like promoters and is dependent on a transcription termination signal present in the leader RNA. *J. Bacteriol.* **184**(11)**:**3034–3043.

Ghosh, P., Bibb, L. A., and Hatfull, G. F. (2008). Two-step site selection for serine-integrase-mediated excision: DNA-directed integrase conformation and central dinucleotide proofreading. *Proc. Natl. Acad. Sci. USA* **105**(9)**:**3238–3243.

Ghosh, P., Kim, A. I., and Hatfull, G. F. (2003). The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of attP and attB. *Mol. Cell.* **12** (5)**:**1101–1111.

Ghosh, P., Pannunzio, N. R., and Hatfull, G. F. (2005). Synapsis in phage Bxb1 integration: Selection mechanism for the correct pair of recombination sites. *J. Mol. Biol.* **349**(2)**:**331–348.

Ghosh, P., Wasil, L. R., and Hatfull, G. F. (2006). Control of phage Bxb1 Excision by a novel recombination directionality factor. *PLoS Biol.* **4**(6)**:**e186.

Gil, F., Catalao, M. J., Moniz-Pereira, J., Leandro, P., McNeil, M., and Pimentel, M. (2008). The lytic cassette of mycobacteriophage Ms6 encodes an enzyme with lipolytic activity. *Microbiology* **154**(Pt 5)**:**1364–1371.

Gil, F., Grzegorzewicz, A. E., Catalao, M. J., Vital, J., McNeil, M. R., and Pimentel, M. (2010). Mycobacteriophage Ms6 LysB specifically targets the outer membrane of *Mycobacterium smegmatis*. *Microbiology* **156**(Pt 5)**:**1497–1504.

Giri, N., Bhowmik, P., Bhattacharya, B., Mitra, M., and Das Gupta, S. K. (2009). The mycobacteriophage D29 gene 65 encodes an early-expressed protein that functions as a structure-specific nuclease. *J. Bacteriol.* **191**(3)**:**959–967.

Author's personal copy

Gomathi, N. S., Sameer, H., Kumar, V., Balaji, S., Dustackeer, V. N., and Narayanan, P. R. (2007). In silico analysis of mycobacteriophage Che12 genome: Characterization of genes required to lysogenise *Mycobacterium tuberculosis*. *Comput. Biol. Chem.* **31**(2)**:**82–91.

Guilhot, C., Gicquel, B., Davies, J., and Martin, C. (1992). Isolation and analysis of IS6120, a new insertion sequence from *Mycobacterium smegmatis*. *Mol. Microbiol.* **6**(1)**:**107–113.

Han, S., Craig, J. A., Putnam, C. D., Carozzi, N. B., and Tainer, J. A. (1999). Evolution and mechanism from structures of an ADP-ribosylating toxin and NAD complex. *Nat. Struct. Biol.* **6**(10)**:**932–936.

Hanauer, D. I., Jacobs-Sera, D., Pedulla, M. L., Cresawn, S. G., Hendrix, R. W., and Hatfull, G. F. (2006). Inquiry learning: Teaching scientific inquiry. *Science* **314**(5807)**:** 1880–1881.

Hassan, S., Mahalingam, V., and Kumar, V. (2009). Synonymous codon usage analysis of thirty two mycobacteriophage genomes. *Adv Bioinformatics*, 1–11.

Hatfull, G. F. (1994). Mycobacteriophage L5: A toolbox for tuberculosis. *ASM News* **60:**255–260.

Hatfull, G. F. (1999). Mycobacteriophages. *In* ''Mycobacteria: Molecular Biology and Virulence'' (C. Ratledge and J. Dale, eds.), pp. 38–58. Chapman and Hall, London.

Hatfull, G. F. (2000). Molecular genetics of mycobacteriophages. *In* ''Molecular Genetics of the Mycobacteria'' (G. F. Hatfull and W. R. Jacobs, Jr., eds.), pp. 37–54. ASM Press, Washington, DC.

Hatfull, G. F. (2004). Mycobacteriophages and tuberculosis. *In* ''tuberculosis'' (K. Eisenach, S. T. Cole, W. R. Jacobs, Jr., and D. McMurray, eds.), pp. 203–218. ASM Press, Washington, DC.

Hatfull, G. F. (2006). Mycobacteriophages. *In* ''The Bacteriophages'' (R. Calendar, ed.), pp. 602–620. Oxford University Press, New York.

Hatfull, G. F. (2008). *Bacteriophage genomics. Curr. Opin. Microbiol.* **11**(5)**:**447–453.

Hatfull, G. F. (2010). Mycobacteriophages: Genes and genomes. *Annu. Rev. Microbiol.* **64:**331–356.

Hatfull, G. F., Barsom, L., Chang, L., Donnelly-Wu, M., Lee, M. H., Levin, M., Nesbit, C., and Sarkis, G. J. (1994). Bacteriophages as tools for vaccine development. *Dev. Biol. Stand.* **82:**43–47.

Hatfull, G. F., Cresawn, S. G., and Hendrix, R. W. (2008). Comparative genomics of the mycobacteriophages: Insights into bacteriophage evolution. *Res. Microbiol.* **159**(5)**:** 332–339.

Hatfull, G. F., and Jacobs, W. R., Jr. (2000). Molecular Genetics of the Mycobacteria. ASM Press, Washington, DC.

Hatfull, G. F., and Jacobs, W. R., Jr. (1994). Mycobacteriophages: Cornerstones of mycobacterial research. *In* ''Tuberculosis: Pathogenesis, Protection and Control'' (B. R. Bloom, ed.), pp. 165–183. ASM, Washington, DC.

Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C. C., Weber, R. J., Patel, M. C., Germane, K. L., Edgar, R. H., Hoyte, N. N., Bowman, C. A., *et al.* (2010). Comparative genomic analysis of 60 mycobacteriophage genomes: Genome clustering, gene acquisition, and gene size. *J. Mol. Biol.* **397**(1)**:**119–143.

Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E., Gonda, R. M., Houtz, J. M., Hryckowian, A. J., Kelchner, V. A., Namburi, S., Pajcini, K. V., *et al.* (2006). Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genet.* **2**(6)**:**e92.

Hatfull, G. F., and Sarkis, G. J. (2006). DNA sequence, structure and gene expression of mycobacteriophage L5: A phage system for mycobacterial genetics. *Mol. Microbiol.* **7**(3)**:** 395–405.

Hayes, C. S., and Keiler, K. C. (2010). Beyond ribosome rescue: tmRNA and co-translational processes. *FEBS Lett.* **584**(2)**:**413–419.

Author's personal copy

Hendrix, R. W. (2002). Bacteriophages: Evolution of the majority. *Theor. Popul. Biol.* **61**(4)**:** 471–480.

Hendrix, R. W. (2003). *Bacteriophage genomics. Curr. Opin. Microbiol.* **6**(5)**:**506–511.

Hendrix, R. W. (2005). Bacteriophage HK97: Assembly of the capsid and evolutionary connections. *Adv. Virus Res.* **64:**1–14.

Hendrix, R. W., Lawrence, J. G., Hatfull, G. F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**(11)**:**504–508.

Hendrix, R. W., Smith, M. C., Burns, R. N., Ford, M. E., and Hatfull, G. F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. USA* **96**(5)**:**2192–2197.

Henry, M., Begley, M., Neve, H., Maher, F., Ross, R. P., McAuliffe, O., Coffey, A., and O'Mahony, J. M. (2010a). Cloning and expression of a mureinolytic enzyme from the mycobacteriophage TM4. *FEMS Microbiol. Lett.* **311**(2)**:**126–132.

Henry, M., O'Sullivan, O., Sleator, R. D., Coffey, A., Ross, R. P., McAuliffe, O., and O'Mahony, J. M. (2010b). In silico analysis of Ardmore, a novel mycobacteriophage isolated from soil. *Gene* **453**(1–2)**:**9–23.

Huff, J., Czyz, A., Landick, R., and Niederweis, M. (2010). Taking phage integration to the next level as a genetic tool for mycobacteria. *Gene* **468**(1–2)**:**8–19.

Jacobs, W. R., Jr. (1992). Advances in mycobacterial genetics: New promises for old diseases. *Immunobiology* **184**(2–3)**:**147–156.

Jacobs, W. R., Jr. (2000). *Mycobacterium tuberculosis*: A once genetically intractable organism. *In* ''Molecular Genetics of the Mycobacteria'' (G. F. Hatfull and W. R. Jacobs, Jr., eds.), pp. 1–16. ASM Press, Washington, DC.

Jacobs, W. R., Jr., Barletta, R. G., Udani, R., Chan, J., Kalkut, G., Sosne, G., Kieser, T., Sarkis, G. J., Hatfull, G. F., and Bloom, B. R. (1993). Rapid assessment of drug susceptibilities of *Mycobacterium tuberculosis* by means of luciferase reporter phages. *Science* **260**(5109)**:**819–822.

Jacobs, W. R., Jr., Kalpana, G. V., Cirillo, J. D., Pascopella, L., Snapper, S. B., Udani, R. A., Jones, W., Barletta, R. G., and Bloom, B. R. (1991). *Genetic systems for mycobacteria. Methods Enzymol.* **204:**537–555.

Jacobs, W. R., Jr., Snapper, S. B., Tuckman, M., and Bloom, B. R. (1989). *Mycobacteriophage vector systems. Rev. Infect. Dis.* **11**(Suppl. 2)**:**S404–S410.

Jacobs, W. R., Jr., Tuckman, M., and Bloom, B. R. (1987). Introduction of foreign DNA into mycobacteria using a shuttle phasmid. *Nature* **327**(6122)**:**532–535.

Jain, S., and Hatfull, G. F. (2000). Transcriptional regulation and immunity in mycobacteriophage Bxb1. *Mol. Microbiol.* **38**(5)**:**971–985.

Johnson, J. E. (2010). Virus particle maturation: Insights into elegantly programmed nanomachines. *Curr. Opin. Struct. Biol.* **20**(2)**:**210–216.

Jones, W. D., Jr. (1975). Phage typing report of 125 strains of ''Mycobacterium tuberculosis. *Ann. Sclavo* **17**(4)**:**599–604.

Kalpana, G. V., Bloom, B. R., and Jacobs, W. R., Jr. (1991). Insertional mutagenesis and illegitimate recombination in mycobacteria. *Proc. Natl. Acad. Sci. USA* **88**(12)**:**5433–5437.

Khoo, K. H., Suzuki, R., Dell, A., Morris, H. R., McNeil, M. R., Brennan, P. J., and Besra, G. S. (1996). Chemistry of the lyxose-containing mycobacteriophage receptors of *Mycobacterium phlei/Mycobacterium smegmatis*. *Biochemistry* **35**(36)**:**11812–11819.

Kim, A. I., Ghosh, P., Aaron, M. A., Bibb, L. A., Jain, S., and Hatfull, G. F. (2003). Mycobacteriophage Bxb1 integrates into the *Mycobacterium smegmatis* groEL1 gene. *Mol. Microbiol.* **50**(2)**:**463–473.

Kleckner, N., Bender, J., and Gottesman, S. (1991). Uses of transposons with emphasis on Tn10. *Methods Enzymol.* **204:**139–180.

Koz'min-Sokolov, B. N., and Vabilin, (1975). Effect of mycobacteriophages on the course of experimental tuberculosis in albino mice. *Probl. Tuberk* **4:**75–79.

Author's personal copy

Krisch, H. M., and Comeau, A. M. (2008). The immense journey of bacteriophage T4: From d'Herelle to Delbruck and then to Darwin and beyond. *Res. Microbiol.* **159**(5)**:**314–324.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**(8)**:**1026–1028.

Kumar, V., Loganathan, P., Sivaramakrishnan, G., Kriakov, J., Dusthakeer, A., Subramanyam, B., Chan, J., Jacobs, W. R., Jr., and Paranji Rama, N. (2008). Characterization of temperate phage Che12 and construction of a new tool for diagnosis of tuberculosis. *Tuberculosis (Edinb.)* **88**(6)**:**616–623.

Kunisawa, T. (2000). Functional role of mycobacteriophage transfer RNAs. *J. Theor. Biol.* **205** (1)**:**167–170.

Lai, X., Weng, J., Zhang, X., Shi, W., Zhao, J., and Wang, H. (2006). MSTF: A domain involved in bacterial metallopeptidases and surface proteins, mycobacteriophage tape-measure proteins and fungal proteins. *FEMS Microbiol. Lett.* **258**(1)**:**78–82.

Lamichhane, G., Zignol, M., Blades, N. J., Geiman, D. E., Dougherty, A., Grosset, J., Broman, K. W., and Bishai, W. R. (2003). A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: Application to Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. USA* **100**(12)**:**7213–7218.

Lawrence, J. G., Hatfull, G. F., and Hendrix, R. W. (2002). Imbroglios of viral taxonomy: Genetic exchange and failings of phenetic approaches. *J. Bacteriol.* **184**(17)**:**4891–4905.

Lee, M. H., and Hatfull, G. F. (1993). Mycobacteriophage L5 integrase-mediated site-specific integration in vitro. *J. Bacteriol.* **175**(21)**:**6836–6841.

Lee, M. H., Pascopella, L., Jacobs, W. R., Jr., and Hatfull, G. F. (1991). Site-specific integration of mycobacteriophage L5: Integration-proficient vectors for *Mycobacterium smegmatis*, *Mycobacterium tuberculosis*, and bacille Calmette-Guerin. *Proc. Natl. Acad. Sci. USA* **88** (8)**:**3111–3115.

Lee, S., Kriakov, J., Vilcheze, C., Dai, Z., Hatfull, G. F., and Jacobs, W. R., Jr. (2004). Bxz1, a new generalized transducing phage for mycobacteria. *FEMS Microbiol. Lett.* **241**(2)**:** 271–276.

Lehnherr, H., Maguin, E., Jafri, S., and Yarmolinsky, M. B. (1993). Plasmid addiction genes of bacteriophage P1: doc, which causes cell death on curing of prophage, and phd, which prevents host death when prophage is retained. *J. Mol. Biol.* **233**(3)**:**414–428.

Levin, M. E., Hendrix, R. W., and Casjens, S. R. (1993). A programmed translational frameshift is required for the synthesis of a bacteriophage lambda tail assembly protein. *J. Mol. Biol.* **234**(1)**:**124–139.

Lewis, J. A., and Hatfull, G. F. (2000). Identification and characterization of mycobacteriophage L5 excisionase. *Mol. Microbiol.* **35**(2)**:**350–360.

Lewis, J. A., and Hatfull, G. F. (2001). Control of directionality in integrase-mediated recombination: Examination of recombination directionality factors (RDFs) including Xis and Cox proteins. *Nucleic Acids Res.* **29**(11)**:**2205–2216.

Lewis, J. A., and Hatfull, G. F. (2003). Control of directionality in L5 integrase-mediated site-specific recombination. *J. Mol. Biol.* **326**(3)**:**805–821.

Lima-Mendez, G., Toussaint, A., and Leplae, R. (2007). Analysis of the phage sequence space: The benefit of structured information. *Virology* **365**(2)**:**241–249.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* **25**(4)**:**762–777.

Marinelli, L. J., Piuri, M., Swigonova, Z., Balachandran, A., Oldfield, L. M., van Kessel, J. C., and Hatfull, G. F. (2008). BRED: A simple and powerful tool for constructing mutant and recombinant bacteriophage genomes. *PLoS One* **3**(12)**:**e3957.

Martin, C., Mazodier, P., Mediola, M. V., Gicquel, B., Smokvina, T., Thompson, C. J., and Davies, J. (1991). Site-specific integration of the Streptomyces plasmid pSAM2 in *Mycobacterium smegmatis*. *Mol. Microbiol.* **5**(10)**:**2499–2502.

Author's personal copy

Martinsohn, J. T., Radman, M., and Petit, M. A. (2008). The lambda red proteins promote efficient recombination between diverged sequences: Implications for bacteriophage genome mosaicism. *PLoS Genet.* **4**(5)**:**e1000065.

McNerney, R. (1999). TB: The return of the phage: A review of fifty years of mycobacteriophage research. *Int. J. Tuberc. Lung Dis.* **3**(3)**:**179–184.

McNerney, R., Kiepiela, P., Bishop, K. S., Nye, P. M., and Stoker, N. G. (2000). Rapid screening of Mycobacterium tuberculosis for susceptibility to rifampicin and streptomycin. *Int. J. Tuberc. Lung Dis.* **4**(1)**:**69–75.

McNerney, R., and Traore, H. (2005). Mycobacteriophage and their application to disease control. *J. Appl. Microbiol* **99**(2)**:**223–233.

Mediavilla, J., Jain, S., Kriakov, J., Ford, M. E., Duda, R. L., Jacobs, W. R., Jr., Hendrix, R. W., and Hatfull, G. F. (2000). Genome organization and characterization of mycobacteriophage Bxb1. *Mol. Microbiol.* **38**(5)**:**955–970.

Mizuguchi, Y. (1984). Mycobacteriophages. *In* ''The Mycobacteria: A Sourcebook'' (G. P. Kubica and L. G. Wayne, eds.), Vol. Part A, pp. 641–662. Marcel Dekker, New York.

Morris, P., Marinelli, L. J., Jacobs-Sera, D., Hendrix, R. W., and Hatfull, G. F. (2008). Genomic characterization of mycobacteriophage Giles: Evidence for phage acquisition of host DNA by illegitimate recombination. *J. Bacteriol.* **190**(6)**:**2172–2182.

Murphy, K. C. (1998). Use of bacteriophage lambda recombination functions to promote gene replacement in *Escherichia coli*. *J. Bacteriol.* **180**(8)**:**2063–2071.

Murry, J., Sassetti, C. M., Moreira, J., Lane, J., and Rubin, E. J. (2005). A new site-specific integration system for mycobacteria. *Tuberculosis (Edinb.)* **85**(5–6)**:**317–323.

Nesbit, C. E., Levin, M. E., Donnelly-Wu, M. K., and Hatfull, G. F. (1995). Transcriptional regulation of repressor synthesis in mycobacteriophage L5. *Mol. Microbiol.* **17**(6)**:** 1045–1056.

Ojha, A., Anand, M., Bhatt, A., Kremer, L., Jacobs, W. R., Jr., and Hatfull, G. F. (2005). GroEL1: A dedicated chaperone involved in mycolic acid biosynthesis during biofilm formation in mycobacteria. *Cell* **123**(5)**:**861–873.

Pai, M., Kalantri, S., Pascopella, L., Riley, L. W., and Reingold, A. L. (2005). Bacteriophage-based assays for the rapid detection of rifampicin resistance in Mycobacterium tuberculosis: A meta-analysis. *J. Infect.* **51**(3)**:**175–187.

Parish, T., Lewis, J., and Stoker, N. G. (2001). Use of the mycobacteriophage L5 excisionase in *Mycobacterium tuberculosis* to demonstrate gene essentiality. *Tuberculosis (Edinb.)* **81** (5–6)**:**359–364.

Pashley, C. A., and Parish, T. (2003). Efficient switching of mycobacteriophage L5-based integrating plasmids in *Mycobacterium tuberculosis*. *FEMS Microbiol. Lett.* **229**(2)**:**211–215.

Payne, K., Sun, Q., Sacchettini, J., and Hatfull, G. F. (2009). Mycobacteriophage Lysin B is a novel mycolylarabinogalactan esterase. *Mol. Microbiol.* **73**(3)**:**367–381.

Pearson, R. E., Jurgensen, S., Sarkis, G. J., Hatfull, G. F., and Jacobs, W. R., Jr. (1996). Construction of D29 shuttle phasmids and luciferase reporter phages for detection of mycobacteria. *Gene* **183**(1–2)**:**129–136.

Pedulla, M. L., Ford, M. E., Houtz, J. M., Karthikeyan, T., Wadsworth, C., Lewis, J. A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N. R., Brucker, W., Kumar, V., *et al.* (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**(2)**:**171–182.

Pedulla, M. L., and Hatfull, G. F. (1998). Characterization of the mIHF gene of *Mycobacterium smegmatis*. *J. Bacteriol.* **180**(20)**:**5473–5477.

Pedulla, M. L., Lee, M. H., Lever, D. C., and Hatfull, G. F. (1996). A novel host factor for integration of mycobacteriophage L5. *Proc. Natl. Acad. Sci. USA* **93**(26)**:**15411–15416.

Pell, L. G., Gasmi-Seabrook, G. M., Morais, M., Neudecker, P., Kanelis, V., Bona, D., Donaldson, L. W., Edwards, A. M., Howell, P. L., Davidson, A. R., and Maxwell, K. L. (2010). The solution structure of the C-terminal Ig-like domain of the bacteriophage lambda tail tube protein. *J. Mol. Biol.* **403**(3)**:**468–479.

Author's personal copy

Pell, L. G., Kanelis, V., Donaldson, L. W., Howell, P. L., and Davidson, A. R. (2009). The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. *Proc. Natl. Acad. Sci. USA* **106** (11)**:**4160–4165.

Peña, C. E., Kahlenberg, J. M., and Hatfull, G. F. (1998). The role of supercoiling in mycobacteriophage L5 integrative recombination. *Nucleic Acids Res.* **26**(17)**:**4012–4018.

Peña, C. E., Kahlenberg, J. M., and Hatfull, G. F. (2000). Assembly and activation of site-specific recombination complexes. *Proc. Natl. Acad. Sci. USA* **97**(14)**:**7760–7765.

Peña, C. E., Lee, M. H., Pedulla, M. L., and Hatfull, G. F. (1997). Characterization of the mycobacteriophage L5 attachment site, attP. *J. Mol. Biol.* **266**(1)**:**76–92.

Peña, C. E., Stoner, J. E., and Hatfull, G. F. (1996). Positions of strand exchange in mycobacteriophage L5 integration and characterization of the attB site. *J. Bacteriol.* **178** (18)**:**5533–5536.

Pham, T. T., Jacobs-Sera, D., Pedulla, M. L., Hendrix, R. W., and Hatfull, G. F. (2007). Comparative genomic analysis of mycobacteriophage Tweety: Evolutionary insights and construction of compatible site-specific integration vectors for mycobacteria. *Microbiology* **153**(Pt 8)**:**2711–2723.

Pitcher, R. S., Brissett, N. C., and Doherty, A. J. (2007). Nonhomologous end-joining in bacteria: A microbial perspective. *Annu. Rev. Microbiol.* **61:**259–282.

Pitcher, R. S., Tonkin, L. M., Daley, J. M., Palmbos, P. L., Green, A. J., Velting, T. L., Brzostek, A., Korycka-Machala, M., Cresawn, S., Dziadek, J., Hatfull, G. F., Wilson, T. E., *et al.* (2006). Mycobacteriophage exploit NHEJ to facilitate genome circularization. *Mol. Cell* **23**(5)**:**743–748.

Pitcher, R. S., Wilson, T. E., and Doherty, A. J. (2005). New insights into NHEJ repair processes in prokaryotes. *Cell Cycle* **4**(5)**:**675–678.

Piuri, M., and Hatfull, G. F. (2006). A peptidoglycan hydrolase motif within the mycobacteriophage TM4 tape measure protein promotes efficient infection of stationary phase cells. *Mol. Microbiol.* **62**(6)**:**1569–1585.

Piuri, M., Jacobs, W. R., Jr., and Hatfull, G. F. (2009). Fluoromycobacteriophages for rapid, specific, and sensitive antibiotic susceptibility testing of *Mycobacterium tuberculosis*. *PLoS ONE* **4**(3)**:**e4870.

Popa, M. P., McKelvey, T. A., Hempel, J., and Hendrix, R. W. (1991). Bacteriophage HK97 structure: Wholesale covalent cross-linking between the major head shell subunits. *J. Virol.* **65**(6)**:**3227–3237.

Pope, W. H., Jacobs-Sera, D., Russell, D. A., Peebles, C. L., Al-Atrache, Z., Alcoser, T. A., Alexander, L. M., Alfano, M. B., Alford, S. T., Amy, N. E., Anderson, M. D., Anderson, A. G., *et al.* (2011). Expanding the diversity of mycobacteriophages: Insights into genome architecture and evolution. *PLoS One* **6**(1)**:**e16329.

Raj, C. V., and Ramakrishnan, T. (1970). Transduction in *Mycobacterium smegmatis*. *Nature* **228** (268)**:**280–281.

Ramesh, G. R., and Gopinathan, K. P. (1994). Structural proteins of mycobacteriophage I3: Cloning, expression and sequence analysis of a gene encoding a 70-kDa structural protein. *Gene* **143**(1)**:**95–100.

Redmond, W. B., and Ward, D. M. (1966). Media and methods for phage-typing mycobacteria. *Bull. World Health Organ.* **35**(4)**:**563–568.

Rubin, E. J., Akerley, B. J., Novik, V. N., Lampe, D. J., Husson, R. N., and Mekalanos, J. J. (1999). In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria. *Proc. Natl. Acad. Sci. USA* **96**(4)**:**1645–1650.

Rybniker, J., Kramme, S., and Small, P. L. (2006). Host range of 14 mycobacteriophages in *Mycobacterium ulcerans* and seven other mycobacteria including *Mycobacterium tuberculosis*: Application for identification and susceptibility testing. *J. Med. Microbiol.* **55**(Pt 1)**:**37–42.

Author's personal copy

Rybniker, J., Nowag, A., van Gumpel, E., Nissen, N., Robinson, N., Plum, G., and Hartmann, P. (2010). Insights into the function of the WhiB-like protein of mycobacteriophage TM4: A transcriptional inhibitor of WhiB2. *Mol. Microbiol.* **77**(3)**:**642–657.

Rybniker, J., Plum, G., Robinson, N., Small, P. L., and Hartmann, P. (2008). Identification of three cytotoxic early proteins of mycobacteriophage L5 leading to growth inhibition in *Mycobacterium smegmatis*. *Microbiology* **154**(Pt 8)**:**2304–2314.

Sahu, K., Gupta, S. K., and Ghosh, T. C. (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. (S. Sau, ed.), *J. Biochem. Mol. Biol.* **37**(4)**:**487–492.

Sampson, T., Broussard, G. W., Marinelli, L. J., Jacobs-Sera, D., Ray, M., Ko, C. C., Russell, D., Hendrix, R. W., and Hatfull, G. F. (2009). Mycobacteriophages BPs, Angel and Halo: Comparative genomics reveals a novel class of ultra-small mobile genetic elements. *Microbiology* **155**(Pt 9)**:**2962–2977.

Sarkis, G. J., Jacobs, W. R., Jr., and Hatfull, G. F. (1995). L5 luciferase reporter mycobacteriophages: A sensitive tool for the detection and assay of live mycobacteria. *Mol. Microbiol.* **15**(6)**:**1055–1067.

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. USA* **98**(22)**:**12712–12717.

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* **48**(1)**:**77–84.

Sau, S., Chattoraj, P., Ganguly, T., Lee, C. Y., and Mandal, N. C. (2004). Cloning and sequencing analysis of the repressor gene of temperate mycobacteriophage L1. *J. Biochem. Mol. Biol.* **37**(2)**:**254–259.

Saviola, B., and Bishai, W. R. (2004). Method to integrate multiple plasmids into the mycobacterial chromosome. *Nucleic Acids Res.* **32**(1)**:**e11.

Scollard, D. M., Adams, L. B., Gillis, T. P., Krahenbuhl, J. L., Truman, R. W., and Williams, D. L. (2006). The continuing challenges of leprosy. *Clin. Microbiol. Rev.* **19**(2)**:**338–381.

Seoane, A., Navas, J., and Garcia Lobo, J. M. (1997). Targets for pSAM2 integrase-mediated site-specific integration in the *Mycobacterium smegmatis* chromosome. *Microbiology* **143**(Pt 10)**:**3375–3380.

Sivanathan, V., Allen, M. D., de Bekker, C., Baker, R., Arciszewska, L. K., Freund, S. M., Bycroft, M., Lowe, J., and Sherratt, D. J. (2006). The FtsK gamma domain directs oriented DNA translocation by interacting with KOPS. *Nat. Struct. Mol. Biol.* **13**(11)**:**965–972.

Smith, M. C., and Thorpe, H. M. (2002). Diversity in the serine recombinases. *Mol. Microbiol.* **44**(2)**:**299–307.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**(7)**:**951–960.

Springer, B., Sander, P., Sedlacek, L., Ellrott, K., and Bottger, E. C. (2001). Instability and site-specific excision of integration-proficient mycobacteriophage L5 plasmids: Development of stably maintained integrative vectors. *Int. J. Med. Microbiol.* **290**(8)**:**669–675.

Stella, E. J., de la Iglesia, A. I., and Morbidoni, H. R. (2009). Mycobacteriophages as versatile tools for genetic manipulation of mycobacteria and development of simple methods for diagnosis of mycobacterial diseases. *Rev. Argent Microbiol.* **41**(1)**:**45–55.

Stewart, C. R., Casjens, S. R., Cresawn, S. G., Houtz, J. M., Smith, A. L., Ford, M. E., Peebles, C. L., Hatfull, G. F., Hendrix, R. W., Huang, W. M., and Pedulla, M. L. (2009). The genome of *Bacillus subtilis* bacteriophage SPO1. *J. Mol. Biol.* **388**(1)**:**48–70.

Sula, L., Sulova, J., and Stolcpartova, M. (1981). Therapy of experimental tuberculosis in guinea pigs with mycobacterial phages DS-6A, GR-21 T, My-327. *Czech. Med.* **4**(4)**:**209–214.

Susskind, M. M., and Botstein, D. (1978). Molecular genetics of bacteriophage P22. *Microbiol. Rev.* **42**(2)**:**385–413.

Author's personal copy

Timme, T. L., and Brennan, P. J. (1984). Induction of bacteriophage from members of the *Mycobacterium avium*, *Mycobacterium intracellulare*. *Mycobacterium scrofulaceum* serocomplex. *J. Gen. Microbiol.* **130**(Pt 8)**:**2059–2066.

Tori, K., Dassa, B., Johnson, M. A., Southworth, M. W., Brace, L. E., Ishino, Y., Pietrokovski, S., and Perler, F. B. (2009). Splicing of the mycobacteriophage Bethlehem DnaB intein: Identification of a new mechanistic class of inteins that contain an obligate block F nucleophile. *J. Biol. Chem.* **285**(4)**:**2515–2526.

Tyler, J. S., Mills, M. J., and Friedman, D. I. (2004). The operator and early promoter region of the Shiga toxin type 2-encoding bacteriophage 933W and control of toxin expression. *J. Bacteriol.* **186**(22)**:**7670–7679.

van Kessel, J. C., and Hatfull, G. F. (2007). *Recombineering in Mycobacterium tuberculosis. Nat. Methods* **4**(2)**:**147–152.

van Kessel, J. C., and Hatfull, G. F. (2008a). Efficient point mutagenesis in mycobacteria using single-stranded DNA recombineering: Characterization of antimycobacterial drug targets. *Mol. Microbiol.* **67**(5)**:**1094–1107.

van Kessel, J. C., and Hatfull, G. F. (2008b). *Mycobacterial recombineering. Methods Mol. Biol.* **435:**203–215.

van Kessel, J. C., Marinelli, L. J., and Hatfull, G. F. (2008). Recombineering mycobacteria and their phages. *Nat. Rev. Microbiol.* **6**(11)**:**851–857.

Vultos, T. D., Mederle, I., Abadie, V., Pimentel, M., Moniz-Pereira, J., Gicquel, B., Reyrat, J. M., and Winter, N. (2006). Modification of the mycobacteriophage Ms6 attP core allows the integration of multiple vectors into different tRNAala T-loops in slow- and fast-growing mycobacteria. *BMC Mol. Biol.* **7:**47.

Watterson, S. A., Wilson, S. M., Yates, M. D., and Drobniewski, F. A. (1998). Comparison of three molecular assays for rapid detection of rifampin resistance in *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* **36**(7)**:**1969–1973.

Wikoff, W. R., Liljas, L., Duda, R. L., Tsuruta, H., Hendrix, R. W., and Johnson, J. E. (2000). Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289** (5487)**:**2129–2133.

Wilson, S. M., al-Suwaidi, Z., McNerney, R., Porter, J., and Drobniewski, F. (1997). Evaluation of a new rapid bacteriophage-based method for the drug susceptibility testing of *Mycobacterium tuberculosis*. *Nat. Med.* **3**(4)**:**465–468.

Xu, J., Hendrix, R. W., and Duda, R. L. (2004). Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell.* **16**(1)**:**11–21.

Zhang, Y., Buchholz, F., Muyrers, J. P., and Stewart, A. F. (1998). A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.* **20**(2)**:**123–128.

Zhu, H., Yin, S., and Shuman, S. (2004). Characterization of polynucleotide kinase/phosphatase enzymes from Mycobacteriophages omega and Cjw1 and vibriophage KVP40. *J. Biol. Chem.* **279**(25)**:**26358–26369.